MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 5, 898–929

Ideal Point or Dominance Process? Unfolding Tree Approaches to Likert Scale Data with Multi-Process Models

Biao Zeng[a] , Hongbo Wen[a], and Minjeong Jeon[b]

[a]Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China; [b]Department of Education, University of California, Los Angeles, LA, USA

ABSTRACT

This study introduces a new multi-process analytical framework based on the ideal point assumption for analyzing Likert scale data with three newly developed Unfolding Tree (UTree) models. Through simulations, we tested the performance of proposed models and existing Item Response Tree (IRTree) models across various conditions. Subsequently, empirical data were utilized to analyze and compare the UTree models relative to IRTree models, exploring respondents' decision-making processes and underlying latent traits. Simulation results showed that fit indices could effectively discern the correct model underlying the data. When the correct model was employed, both IRTree and UTree accurately retrieved item and individual parameters, with the recovery precision improving as the number of items and sample size increased. Conversely, when an incorrect model was utilized, the misspecified model consistently returned biased results in estimating individual parameters, which was pronounced when the respondents followed an ideal point response process. Empirical findings highlight that respondents' decisions align with the ideal point process rather than the dominance process. The respondents' choices of extreme response options are more driven by target traits than by extreme response style. Furthermore, evidence indicates the presence of two distinct but moderately correlated target traits throughout the different decision stages.

KEYWORDS

Item response tree model; unfolding tree model; dominance process; ideal point process; likert scale; extreme response style; latent traits

Introduction

Background

Originating from early 20th-century psychometrics, the Likert scale has become an indispensable tool in both psychology and education fields. Whether in measurement scales of psychology and education or large-scale international educational quality assessment projects such as NAEP, PISA, and TIMSS, a multitude of researchers and programs have employed the Likert scale. By presenting a situational description of a certain subject, respondents are asked to choose from several options the one that is most suitable to their situations, thereby measuring their target traits (Likert, 1932).

However, numerous researchers have identified that respondents frequently exhibit response styles (therefore, response bias) when interacting with these scales (Austin et al., 2006; Wetzel et al., 2013; Zeng et al., 2020). This pertains to the respondents' consistent inclination to favor certain items based on criteria divergent from the intended measurement traits (Paulhus, 1991). Among these, the extreme response style (ERS) stands out as the most prevalent, representing a systematic proclivity to choose extreme options, such as "strongly agree" or "strongly disagree", irrespective of the test content (Van Vaerenbergh & Thomas, 2013). Such preferences may either underestimate or overestimate the actual levels of respondents' target traits (Kim & Bolt, 2021). Concurrently, these styles have the potential to alter the intrinsic dimensions of the scale (Arce-Ferrer & Ketterer, 2003; Baumgartner & Steenkamp, 2001; Moors, 2003), as a result, reducing the reliability of the questionnaires (Clarke, 2001; Hui & Triandis, 1989). This, in turn, can

negatively influence the scale's validity (Arce-Ferrer & Ketterer, 2003; Zeng et al., 2020; Zeng et al., 2024).

To identify the extreme response styles effectively in Likert scale measurements, researchers have employed multiple analytical methods, including descriptive statistics (Greenleaf, 1992), Confirmatory Factor Analysis (Billiet & McClendon, 2000; Zeng et al., 2020), and Latent Class Analysis (Moors, 2003). In recent years, scholars have introduced assessment methods for response styles based on Item Response Theory (IRT), such as the Multi-dimensional Nominal Response Model (Johnson & Bolt, 2010) and the Mixture Partial Credit Model (Austin et al., 2006). Compared to traditional measurement models, IRT models enable the estimation of response style biases at person levels. However, while utilizing these IRT models, it becomes challenging to observe the cognitive decision-making processes that respondents might undergo at various stages, making it difficult to effectively separate response styles from target traits. To measure the cognitive processing of respondents during their responses and more effectively discern response style biases, a multi-process tree-structured IRT model, called Item Response Tree (IRTree) model subsequently emerged.

Limitations of previous research

There are several gaps in existing research on Likert scale data analysis under multi-process IRT framework. Firstly, previous research employs IRTree models that are based on the assumption of the dominance response process to analyze Likert scales. However, numerous studies found that when respondents answer the Likert scales, they often follow an ideal point decision-making process. In such cases, utilizing IRT models based on the dominance response can lead to significant biases when estimating respondent abilities (Fang, 2020). Conversely, employing unfolding models may effectively address this issue (Chernyshenko et al., 2007; Guo et al., 2006; Stark et al., 2006; Tay et al., 2009). However, multi-process decision analysis models are rarely founded on unfolding models in the literature (Jin et al., 2022), making it challenging to delve into whether and how respondents follow an ideal point process in their multi-stage decision-making when answering Likert scales. Therefore, it's necessary to develop Unfolding Tree (UTree) models anchored in the unfolding framework, and further discern whether respondents are basing their decisions to agree or disagree with item statements on a dominance or an ideal point process.

Furthermore, once the latent decision-making process is determined, many studies typically assume that respondents choose extreme phrasing options in the second stage based on extreme response styles (Böckenholt, 2017; Jin et al., 2022; LaHuis et al., 2019; Park & Wu, 2019). However, this assumption may not be universally valid, for example, Kim and Bolt (2021) found using data from the Trends in International Mathematics and Science Study 2015 that only 32% of students responded to the questionnaire based on extreme response style, whereas a larger proportion, 68%, responded based on target traits. Although a few researchers have begun to explore the use of unfolding models within IRTree frameworks, these attempts have primarily focused on singular ERS-based models (Jin et al., 2022). This overlooks that different nodes might be based on target traits (Kim & Bolt, 2021). Additionally, the reliance on single-parameter settings may limit the models' flexibility and interpretability (Li et al., 2025). Moreover, the performance of this type of model under various measurement conditions, such as different test lengths, trait correlations, and small sample sizes, has not been thoroughly investigated, and its efficacy remains to be validated (Jin et al., 2022). In reality, respondents might select extreme options like "strongly agree" because their trait level is very close to the item location. Misusing models built on ERS in the second decision-making stage, by misinterpreting the target trait as extreme response styles, is conceptually incorrect. The estimated parameters fundamentally lose their real-world relevance in such instances. Hence, there is a need to consider ordinal models based on target traits for the second decision-making stage. These models can help us understand whether respondents choose extreme descriptive options based on ERS or the target trait.

Continuing with this scenario, if respondents base their choices in the second decision-making stage on the target trait, the analysis would require a more detailed examination. The respondents might rely on the same single latent target trait in both stages, or they could rely on two different types of target traits in each stage. For instance, in the first stage, they might choose "agree" or "disagree" based on the agreement trait, and in the second stage, they might select "strongly agree/disagree" or "somewhat agree/disagree" based on the degree of agreement (Jeon et al., 2017). If respondents possess two different target traits, relying on a model that assumes a single trait would substantially obscure the diversity of respondents' latent traits, resulting in misleading interpretations of their actual decision-making process

and the underlying traits. This can cause substantial errors in the interpretation of the results, greatly reducing the validity of conclusions. Therefore, it is crucial to determine whether the decisions across the two stages rely on the same target trait or two slightly different traits. However, the answer to this question is still unclear, especially within the context of the ideal point response process. To address this uncertainty, we need to consider models that either hypothesize the existence of a single target trait or propose two distinct target traits. A subsequent comparison of these models is paramount to determine whether respondents' decisions across the stages stem from a single target trait or two distinct ones.

Moreover, once models that align with respondents' decision-making processes and the underlying latent trait types are established, questions arise about the stability and accuracy of the estimation results from these multi-process models under different conditions, especially with the new Unfolding Tree model. The uncertainties surrounding these factors present significant challenges and difficulties in using the models correctly (Jin et al., 2022). A major concern is the unknown impact on individual latent trait estimations when an incorrect model is used. Gaining insight into the extent and nature of these effects is crucial, emphasizing the importance of employing the appropriate model and ensuring its rational and efficient application. Lastly, the majority of research on multi-process IRT models predominantly adopts Markov Chain Monte Carlo estimation methods. This approach can be time-intensive, potentially compromising the models' practicality and accessibility.

Objectives of this study

Building on the foundation of existing IRTree models and integrating the unfolding model, this study aims to establish a set of multi-process IRT models using an unfolding approach. Based on this groundwork, we will create three distinct types of Unfolding Tree models. Our initial objective is to evaluate the performance of both the traditional IRTree and the newly developed UTree models through rigorous simulation studies, thereby elucidating the conditions, importance, and necessity for the accurate and judicious use of these models. Following the simulations, empirical research will contrast traditional IRTree models with the UTree models. The central aim is to discern the specific traits upon which respondents base their responses to Likert scales and to illuminate the underlying decision-making processes. Ultimately, we seek to determine which multi-process IRT models most closely approximate the actual response patterns of the participants.

To answer the overarching research objectives, this study aims to address the following research questions:

1. How do the IRTree and UTree models perform under different conditions? What factors influence the performance of these models?
2. What are the consequences for estimations when a model misaligned with respondents' true decision-making process is used?
3. When responding to Likert scales in real-world scenarios, do respondents rely on the dominance or the ideal point decision-making process? In the second decision-making stage, when opting for extreme responses, are respondents' choices driven by an extreme response style or by specific target traits? If the latter, are these target traits consistent across varied decision-making stages?

The rest of this paper is organized as follows: Section 2 presents a literature review on the existing IRTree and unfolding models. Section 3 proposes the unfolding approach-based multi-process IRT models. Section 4 details the simulation study, aiming to evaluate the performance of various IRTree and UTree models under diverse conditions and examine the repercussions of employing incorrect models. Section 5 describes the empirical application, wherein the alignment of the models with real-world Likert scale response data is assessed, and the potential response processes and underlying latent traits are investigated. Finally, Section 6 provides a discussion that summarizes the key findings of the research and offers a deeper interpretation and discussion of these insights.

Literature review

Item response tree model

The central premise of the tree-structured IRT model revolves around a tree-like structure, akin to decision trees, which breaks down respondents' answers on the Likert scale into multiple stages. This design captures the sequential or nested multi-stage cognitive decision-making processes of respondents (Böckenholt, 2012; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). Consequently, it is often referred to as the Item Response Tree model or the multi-process IRT Model. The model's most notable strength lies in its ability to

segment response behaviors into various decision processes. By allowing respondents to provide distinct responses at different processing stages, the model facilitates the isolation of response styles from target traits. Consequently, this results in an enhanced accuracy in estimating target traits. Furthermore, beyond its precision, the model is also characterized by its significant flexibility and interpretability (Böckenholt, 2012; Jeon & De Boeck, 2016).

The IRTree model primarily comprises three distinct types: linear, nested, and mixed multi-node IRTree (Jeon & De Boeck, 2016). Among them, the nested model is considered the most suitable IRTree model for analyzing the cognitive decision-making process in Likert scales with an even number of response options for agreement. Researchers predominantly employ this model for the analysis of response styles in Likert scales, including two typical models: the Extreme Response Style IRTree (ERS) Model and the Ordinal IRTree (ORD) Model.

Extreme response style IRTree model

The ERS IRTree model refers to the IRTree model designed to measure extreme response styles. This model categorizes respondents' responses into multiple cognitive decision-making stages. Using a four-point Likert scale as an example, the ERS model can divide respondents' answers into two cognitive decision-making stages: (1) Stage 1: Respondents decide whether they agree or disagree with the statement of the item, representing the level of their attitudes. (2) Stage 2: Respondents choose to either agree or disagree with an extremely phrased statement, such as selecting options like "strongly disagree" or "strongly agree", indicating the intensity of their extreme response style (Böckenholt, 2017; Jeon & De Boeck, 2019a; Jeon & De Boeck, 2019b; Park & Wu, 2019), as illustrated in Figure 1.

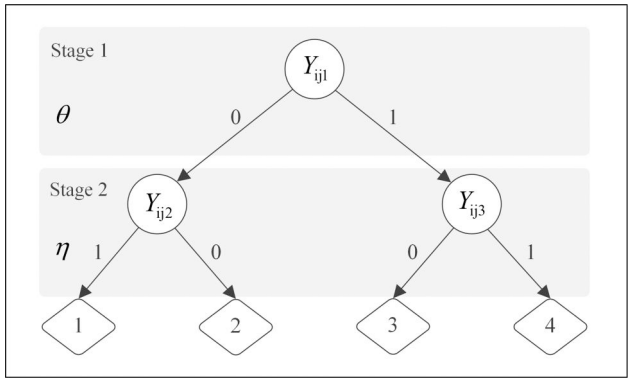

Figure 1. Decision tree diagram of responses on a four-point Likert scale under the ERS IRTree model.

Specifically, we can transform answers from a four-point Likert scale into pseudo-item scores across three distinct decision nodes: Node 1 relates to decision Stage 1, while Nodes 2 and 3 pertain to decision Stage 2. Assuming the response chosen by individual j for item i is represented as Y_{ij1}: In Node 1 (Y_{ij1}), respondents are scored '1′ if they agree with the item's statement and '0′ if they disagree. In Node 2 (Y_{ij2}), among those disagreeing with the statement, a '1′ score is attributed for selecting "strongly disagree" and a '0′ for opting for "somewhat disagree". Conversely, in Node 3 (Y_{ij3}), for those agreeing with the statement, they receive a score of '1′ if they choose "strongly agree" and '0′ if they select "somewhat agree". The scoring method for the four-point Likert scale transformed into pseudo-items is delineated in Table 1.

In Node 1 ($k = 1$), Y_{ij1} equal to 1 indicating agreement with the item statement. Using a two-parameter IRT (2PL) model (Lord, 1952), the probability of individual j agreeing to item i is calculated as:

$$P(Y_{ijk} = 1|\theta_j) = \frac{\exp[\alpha_{ik}(\theta_j - \beta_{ik})]}{1 + \exp[\alpha_{ik}(\theta_j - \beta_{ik})]}. \quad (1)$$

For Nodes 2 and 3 ($k = 2$ or 3), the probability of individual j choosing the extreme phrasing option ("strongly disagree" and "strongly agree") when answering item i is

$$P(Y_{ijk} = 1|\eta_j) = \frac{\exp[\alpha_{ik}(\eta_j - \beta_{ik})]}{1 + \exp[\alpha_{ik}(\eta_j - \beta_{ik})]}. \quad (2)$$

Herein, exp refers to the exponential function with base e. β_{ik} denotes the item difficulty parameter for item i at node k. θ_j represents the target trait level of individual j, while η_j indicates the extreme response style level of individual j. A larger value of θ_j implies that the respondent is more inclined to agree with the item statement and a greater value of η_j suggests a stronger extreme response style. Typically, θ_j and η_j are assumed to jointly follow a standard multivariate normal distribution, where the mean of the distribution is zero and its covariance matrix is the identity matrix.

Ultimately, the probability of individual j choosing each option for item i is the product of the response probabilities across the three nodes, expressed as:

Table 1. Scoring method for the four-point Likert scale transformed into pseudo-items in the ERS IRTree model.

Response (Y_{ij})	Node 1 (Y_{ij1})	Node 2 (Y_{ij2})	Node 3 (Y_{ij3})
1 (Strongly disagree)	0	1	–
2 (Somewhat disagree)	0	0	–
3 (Somewhat agree)	1	–	0
4 (Strongly agree)	1	–	1

$$P(Y_{ij}=1|\theta_j,\eta_j)=P\left(Y_{ij1}=0|\theta_j\right)\times P\left(Y_{ij2}=1|\eta_j\right)$$
$$P(Y_{ij}=2|\theta_j,\eta_j)=P\left(Y_{ij1}=0|\theta_j\right)\times P\left(Y_{ij2}=0|\eta_j\right)$$
$$P(Y_{ij}=3|\theta_j,\eta_j)=P\left(Y_{ij1}=1|\theta_j\right)\times P\left(Y_{ij3}=0|\eta_j\right)$$
$$P(Y_{ij}=4|\theta_j,\eta_j)=P\left(Y_{ij1}=1|\theta_j\right)\times P\left(Y_{ij3}=1|\eta_j\right)$$

It is important to note that the IRTree model is a type of conditional response model, where the probability of an individual ultimately selecting a specific option is a conditional probability. For example, in the ERS model, the probability of an individual selecting a particular option is conditional on the choice made at Node 1 based on θ_j, and the specific choice at Node 2 based on the extreme response style η_j. For instance, the probability of an individual selecting the "strongly disagree" option is the probability of choosing "disagree" at Node 1, under the condition that the individual then enters Node 2 and chooses an extreme response option. This probability can also be expressed as $P(Y_{ij}=1|\theta_j,\eta_j)\ P\left(Y_{ij2}=1|\eta_j,\ \ Y_{ij1}=0,\theta_j\right)$. However, it is important to note that the probability of an individual choosing either "0" or "1" at each node is independent of other node selections. Except for the ERS model, other IRTree models follow a similar logic. Clarifying this point may help in better understanding the inherent logic of this model.

Although the above ERS model has a solid construct, it makes a strong assumption that respondents choose extreme responses in Nodes 2 and 3 due to an extreme response style. In reality, however, respondents might opt for more extreme options because of a particularly high or low level of the target trait. The ERS model in Nodes 2 and 3 only accounts for extreme response style, failing to capture the response process based on the target traits. As a result, researchers have suggested considering the use of the Ordinal IRTree model (Kim & Bolt, 2021).

Ordinal IRTree model

The ORD IRTree model is similar to the ERS IRTree model. However, the ORD model posits that respondents' decisions in Nodes 2 and 3 are based on the measured target trait rather than the extreme response style (Kim & Bolt, 2021). In this manner, the ORD model can more accurately capture instances where respondents, due to a particularly strong or weak target trait, opt for extreme options. Consequently, the scoring method in Node 2 of the ORD model is the exact opposite of the ERS model's method, as detailed in Figure 2.

Building on this, we can derive the transformed pseudo-item scoring as presented in Table 2. In this model, options representing a higher intensity of agreement in Nodes 2 and 3 (i.e., "somewhat disagree" and "strongly agree") are scored as '1′, while the remaining options, representing a lower degree of agreement, are scored as '0′.

In all nodes (k = 1, 2, and 3), the probability of individual j selecting an option representing a higher degree of agreement for item i is given by:

Table 2. Scoring method for the four-point Likert scale transformed into pseudo-items in the ORD IRTree model.

Response (Y_{ij})	Node 1 (Y_{ij1})	Node 2 (Y_{ij2})	Node 3 (Y_{ij3})
1 (Strongly disagree)	0	0	–
2 (Somewhat disagree)	0	1	–
3 (Somewhat agree)	1	–	0
4 (Strongly agree)	1	–	1

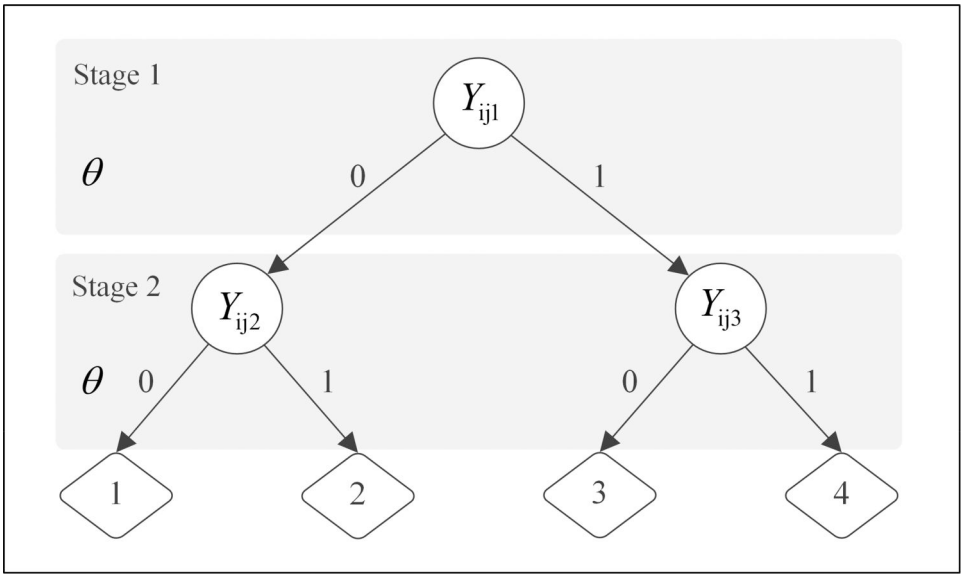

Figure 2. Decision tree diagram of responses on a four-point Likert scale under the ORD IRTree model.

$$P(Y_{ijk}=1|\theta_j)=\frac{\exp[\alpha_{ik}(\theta_j-\beta_{ik})]}{1+\exp[\alpha_{ik}(\theta_j-\beta_{ik})]}. \quad (4)$$

Ultimately, the probability of individual j choosing a specific option for item i is:

$$\begin{aligned}
P(Y_{ij}=1|\theta_j)&=P(Y_{ij1}=0|\theta_j)\times P(Y_{ij2}=0|\theta_j)\\
P(Y_{ij}=2|\theta_j)&=P(Y_{ij1}=0|\theta_j)\times P(Y_{ij2}=1|\theta_j)\\
P(Y_{ij}=3|\theta_j)&=P(Y_{ij1}=1|\theta_j)\times P(Y_{ij3}=0|\theta_j)\\
P(Y_{ij}=4|\theta_j)&=P(Y_{ij1}=1|\theta_j)\times P(Y_{ij3}=1|\theta_j)
\end{aligned} \quad (5)$$

The ORD model posits that a single latent trait, denoted as 'θ_j' and assumed to follow a standard normal distribution with a mean of zero and a standard deviation of one, governs decision-making across two decision stages. This is a critical departure from the ERS model, which presupposes distinct latent traits (θ_j and η_j) for different decision stages, representing fundamentally different conceptualizations of respondent characteristics.

However, the ORD model's assumption that only one latent trait θ_j exists across two decision stages is a very strong assumption. Previous research found that in the IRTree model, respondents might make decisions at different stages based on different target traits (Jeon et al., 2017). For instance, Stage 1 typically represents the latent trait of the degree of agreement with the item statement, whereas Stage 2 (encompassing both Nodes 2 and 3) represents the 'intensity of agreement' trait. Yet, there currently exists no ORD model that allows for different target traits to be the basis for decisions at different stages. This poses significant difficulties and challenges for researchers aiming to explore and interpret decision-making processes based on varying target traits.

In real-world scenarios where respondents' behaviors across different decision-making stages are based on various target traits, there would be two or more distinct types of target traits. In such cases, relying solely on an ORD model that assumes a single target trait is limited. Firstly, assuming only one target trait, unfortunately, leads to a loss of information about the respondents' target traits. For instance, respondents might go through a two-stage decision-making process when answering items related to attitudes toward purchasing certain products. The first stage might be based on whether the product's price falls within their consumption range and the second stage might involve further determination of purchase intent based on product quality. At this point, there exist two different latent traits based on the judgment of the product's price and quality, corresponding to the attitude of agreement to purchase and the intensity of that attitude. Simplistically attributing respondents' decisions to a single latent attitude trait ignores the actual multi-stage decision-making process and loses the diverse latent trait information.

More importantly, an ORD model assuming a single latent trait could result in a misleading interpretation of what is happening in reality (De Boeck & Partchev, 2012). If we simplistically attribute respondents' attitudes toward purchasing products to a general purchasing willingness, it fails to capture the two distinct target traits based on product price and quality, as well as the interaction between these traits. For example, some products might be expensive but of high quality, and despite the cost, the complementary nature of these factors might still incline respondents to purchase. However, using a general purchasing attitude makes it difficult to explain such phenomena, leading to confusion and challenges in interpreting and analyzing why respondents make such decisions. This results in misleading guidance on the respondents' true decision-making process, which is vastly different from the actual situation.

Nevertheless, if we allow the ORD model to assume different types of target traits at various decision-making nodes, it becomes possible to estimate latent traits based on judgments of product price and quality at respective stages. This approach allows for a more accurate and objective interpretation of the respondents' decision-making process, and better explains the moderate attitudes toward purchasing products that are either high in price but good in quality or low in price but poor in quality, thereby offering a more nuanced understanding of the multi-trait decision-making process of respondents.

Hence, in the present study, we propose an ORD model that allows for different target traits across the two decision-making stages. To differentiate it from the existing ORD model that assumes a single decision-making trait, we name the model that allows for two target traits as the "ORD.2" model, and the model that assumes a single target trait as the "ORD.1" model. While the ORD.2 model aligns with the ORD.1 model in terms of the pseudo-item scoring conversion, their specific calculation model configurations differ.

In Node 1 (k=1), the probability of individual j selecting agreement for item i is given by:

$$P(Y_{ijk}=1|\theta_{j1})=\frac{\exp[\alpha_{ik}(\theta_{j1}-\beta_{ik})]}{1+\exp[\alpha_{ik}(\theta_{j1}-\beta_{ik})]}. \quad (6)$$

In Nodes 2 and 3 (k = 2 or 3), the probability of individual j responding to item i with "somewhat disagree" or "somewhat agree" is:

$$P(Y_{ijk}=1|\theta_{j2})=\frac{\exp[\alpha_{ik}(\theta_{j2}-\beta_{ik})]}{1+\exp[\alpha_{ik}(\theta_{j2}-\beta_{ik})]}. \quad (7)$$

where θ_{j1} represents the agreement with the statement and θ_{j2} represents the intensity of agreement. A larger θ_{j1} suggests that the respondent is more likely to agree with the item statement. A larger θ_{j2} indicates a stronger agreement intensity: respondents in Node 2 are more likely to choose the "somewhat disagree" option, while those in Node 3 are more likely to choose the "strongly agree" option. Typically, θ_{j1} and θ_{j2} are presumed to jointly follow a standard multivariate normal distribution. The probability of individual j selecting option m for item i in the ORD.2 model is similar to that in the ORD.1 model, and is therefore not elaborated further here.

Unfolding model

However, these IRTree models such as ERS or ORD were all constructed based on the IRT models, which hypothesize respondents answer the scales according to a dominant response process. This means that the probability of a positive response from respondents increases monotonically with the level of the target trait, as depicted in the left picture of Figure 3. However, numerous studies found that the response process of participants on Likert attitude scales might not adhere to this dominant response process, but the ideal point response process (right picture of Figure 3) (Chernyshenko et al., 2001; 2007; Roberts et al., 2000; Roberts & Laughlin, 1996).

Using a four-point Likert scale as an example, which requires individuals to choose the option that best reflects their situation based on the item statement.

- I have a moderate interest in learning mathematics.

(1) strongly disagree (2) somewhat disagree (3) somewhat agree (4) strongly agree

The dominant response process posits that as a participant's intrinsic interest in mathematics grows, they are more inclined to select the "strongly agree" option. However, this might not truly mirror the participant's decision-making process. Researchers found that those with a moderate interest in math—especially when their interest aligns directly with the item's description—would usually be most likely to opt for the "strongly agree" option (Chernyshenko et al., 2001; 2007). This presents a conundrum where the IRTree model, grounded in the IRT model's dominant response process, may not be optimal for analyzing Likert scales.

To more accurately capture such decision processes of participants when responding to Likert scales, researchers have advanced the concept of an ideal point response process and the unfolding model grounded in it. Contrary to the dominant response process, the ideal point decision-making process hypothesizes that respondents are more inclined to agree with items that closely align with their target trait level (Chernyshenko et al., 2001; 2007; Coombs, 1950; Roberts et al., 2000; Roberts & Laughlin, 1996; Thurstone, 1928). To elucidate, consider a neutral item: this model postulates that a participant's likelihood of agreement peaks when their target trait level aligns perfectly with the item parameters, termed the "ideal point". This likelihood diminishes when the participant's trait level veers too high or too low from this ideal point (Chernyshenko et al., 2001), as illustrated in the right side of Figure 3.

Using the aforementioned four-point Likert scale for illustration, the unfolding model, grounded in the ideal point process, hypothesizes that when a participant's interest in mathematics aligns perfectly with the item's description (i.e., moderate interest in

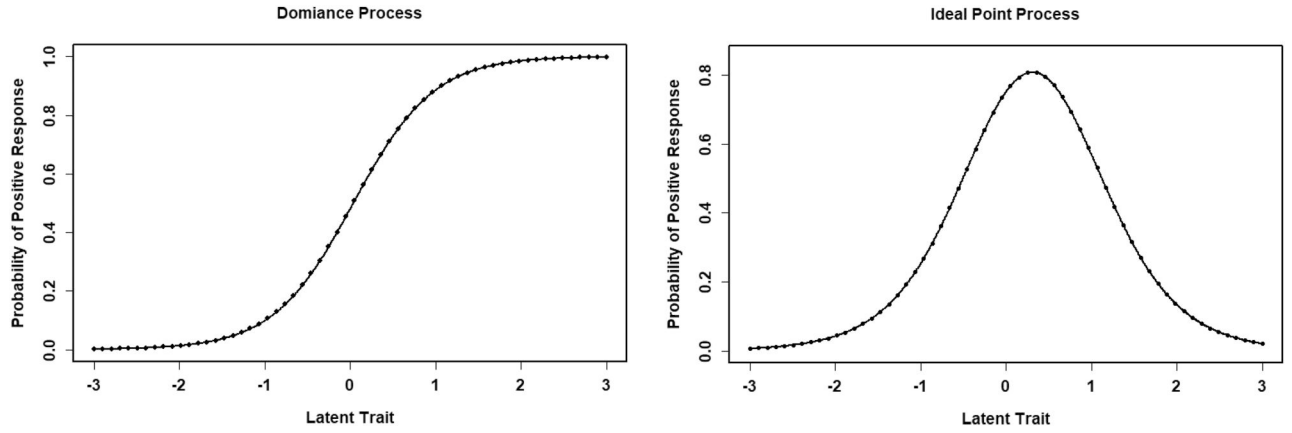

Figure 3. The example of probability of positive response for the IRT model (left) and unfolding model (right).

mathematics), they are most inclined to select the "strongly agree" option. If their interest either increases or decreases significantly, deviating from the central theme of the item (such as higher or lower interest), they are more likely to select the "somewhat agree" option. As the discrepancy between their interest level and the item's theme grows (e.g., very high or very low interest), they tend to choose the "somewhat disagree" option. Finally, when their interest level is in stark contrast to the item's statement (either extremely high or extremely low), they are most likely to choose the "strongly disagree" option. This behavior aligns well with the observed response processes of participants interacting with attitude-oriented Likert scales (Cao et al., 2015; Drasgow et al., 2010; Tay et al., 2009).

Ample studies have underscored the validity of the unfolding model for attitude and personality Likert scales (Cao et al., 2015; Guo et al., 2006; Stark et al., 2006; Tay et al., 2009). For instance, when the unfolding model was employed to analyze the two-point scaled 16PF personality questionnaire, it was found to fit the data more accurately than the dominance IRT model (Stark et al., 2006). Additionally, the ideal point model also exhibits some unique advantages. This model demonstrates psychometric benefits over the dominance model, such as accurately identifying the dimensionality of scales (Tay & Drasgow, 2012). Moreover, it enhances measurement precision and provides more information for individuals with relatively extreme personality traits when neutral items are utilized (Cao et al., 2015; Drasgow et al., 2010). Given this, researchers may consider adopting the unfolding model when constructing an IRTree model. This would lead to the development of an unfolding Tree model. Such a model might be better equipped to capture the ideal point response process of participants' reactions to attitude Likert scales, providing a more precise estimation of both potential response styles and the participants' target traits.

To construct a UTree model, it is essential to select an apt unfolding model with commendable estimation performance. As the field has advanced, numerous unfolding models have been proposed, both for dichotomous and polytomous response data. The primary dichotomous models include the Squared Simple Logistic Model (SSLM) (Andrich, 1988), PARELLA model (Hoijtink, 1991), and Hyperbolic Cosine Model (HCM) (Andrich & Luo, 1993). In contrast, polytomous unfolding models chiefly comprise the Graded Unfolding Model (GUM) (Roberts & Laughlin, 1996) and the Generalized Graded Unfolding Model (GGUM) (Roberts et al., 2000).

The GGUM, in particular, stands out due to its flexibility. It allows for the unconstrained estimation of the discrimination parameter and can be applied to both dichotomous and polytomous data, thus offering an advantage in practice. Numerous studies have also underscored the GGUM model's standout performance in both simulated and real-world scenarios when evaluating personality Likert data (Chernyshenko et al., 2007; Stark et al., 2006), notably with two-point scales (Guo et al., 2006).

Given the flexibility and excellent performance of the GGUM model, the current study intends to use this model to construct the UTree model. Before doing so, let us first introduce the GGUM model. The model categorizes the respondents' reactions into four groups: "strongly agree", "somewhat agree", "somewhat disagree", and "strongly disagree". Among them, "strongly agree" stands alone, while the remaining categories each contain symmetric directions on both the left and right sides. The GGUM centers around the axis $\theta_j - \delta_i = 0$. Respondents whose distances from the central axis are equal will have identical probabilities of selecting the same option.

The GGUM model is formulated as:

$$\mathrm{P}\left(Y_{ij}=\mathrm{y}|\theta_j,\delta_i,\alpha_i,\tau_{is}\right) = \frac{\exp\left\{\alpha_i\left[y\left(\theta_j-\delta_i\right)-\sum_{s=0}^{y}\tau_{is}\right]\right\}+\exp\left\{\alpha_i\left[(M-y)\left(\theta_j-\delta_i\right)-\sum_{s=0}^{y}\tau_{is}\right]\right\}}{\sum_{w=0}^{C}\left\{\exp\left\{\alpha_i\left[w\left(\theta_j-\delta_i\right)-\sum_{s=0}^{w}\tau_{is}\right]\right\}\right\}+\exp\left\{\alpha_i\left[(M-w)\left(\theta_j-\delta_i\right)-\sum_{s=0}^{w}\tau_{is}\right]\right\}}. \quad (8)$$

where Y_{ij} $(0, 1, 2 \ldots C)$ represent the degree of agreement of individual j on the specific options in item i, and the value ranges from 0 to C, where $Y_{ij} = 0$ means strongly disagree, $Y_{ij} = C$ means strongly agree; α_i represents the discrimination of item i; δ_i represents the position of item i on the continuum; θ_j represents the position of person j on the continuum; τ_{is}

represents the threshold limit value of the sth subjective category corresponding to the location of the ith item, $\theta_j - \delta_i = 0$ is the axis of symmetry, and the values on both sides of the axis are the same, $\tau_{i0} = 0$. The probability of agreeing symmetrically mirrors on either side of the axis where $\theta_j - \delta_i = 0$. M is the number of possible response categories, $M = 2C + 1$; w represents the subjective response category.

Given the apparent complexity of this model, Luo (2001) proposed a more streamlined general expression:

$$P(Y_{ij} = 1|\theta_j, \delta_i, \alpha_i, \tau_{is}) = \frac{\psi(\tau_{is})}{\psi\left[\alpha_i(\theta_j - \delta_i)\right] + \psi(\tau_{is})} \quad (9)$$

where the link function $\psi_{(\cdot)}$ can be expressed as:

$$\psi(x_s) = \frac{\cosh\left[\left(\frac{2C+1}{2} + 1 - s\right)x\right]}{\cosh\left[\left(\frac{2C+1}{2} - s\right)x\right]} \quad (10)$$

Considering the simplicity of this expression and its ability to avoid the understanding difficulties caused by complex mathematical formulas, we will use it to represent unfolding models in the following sections.

Proposed model: unfolding tree models

Based on the three types of IRTree models discussed in Section 2, we employ the GGUM model (Roberts et al., 2000), which allows for the free estimation of three parameters ($\delta_i, \alpha_i, \tau_{is}$), to construct three UTree models that represent potential ideal point decision processes. These models are named: ERSUTree, ORDUTree.1, and ORDUTree.2. While we opt for the freely estimated three-parameter GGUM in this study, it is entirely possible to employ a GGUM with a different number of freely estimated parameters or replace it with a different unfolding model, such as the SSLM, PARELLA, HCM, or GUM, as needed. Additionally, although this paper uses a 4-point Likert scale as an example for model construction, the model can be flexibly applied to more granular scoring scales, such as 5, 7, or 9-point Likert scales, thereby enabling the analysis of three or even more nodes of binary outcomes. This highlights the adaptability and flexibility inherent in constructing UTree models.

Currently, Jin et al. (2022) and Li et al. (2025) made preliminary attempts to apply the multi-process model to unfolding models and used it to analyze extreme response styles, finding that individual responses to Likert scales are more likely based on the ideal point process. However, these modeling efforts primarily focused on using the ERS assumption in Nodes 2 and 3, neglecting the possibility that individuals at these nodes might base their responses on target traits (Kim & Bolt, 2021), and some models are based on single-parameter settings (Li et al., 2025), which may limit the model's flexibility. This paper takes a more comprehensive approach by considering the possibility of both ERS and target traits, while allowing the item parameters and traits of different nodes to be freely estimated, thus resulting in the development of ERSUTree, which allows for the free estimation of ERS and θ, as well as ORDUTree.1 and ORDUTree.2, which include both a single trait and two target traits.

Extreme response style unfolding tree model

The ERSUTree model, similar to the ERS model, posits that individuals use a target trait (θ), which represents their attitude toward the item, to decide whether to agree or disagree with an item at Node 1. At Nodes 2 and 3, individuals base their choices on their extreme response style (η), opting for either extreme expression options ("strongly agree" or "strongly disagree"). This model hypothesizes that respondents adhere to the ideal point response process, implying that at Node 1, responses are constructed using the GGUM based on the ideal point process. When a participant's latent trait aligns more closely with the item's position, they are more likely to agree with the item's statement. A schematic representation of the response probability can be seen in Figure 4A.

At Nodes 2 and 3, since participants base their choices on their extreme response style, the probability of a participant opting for an extreme expression option increases as the level of extreme response style increases. This increment follows a monotonically increasing dominance response process. The response probability's schematic representation can be depicted as Figure 4B. Subsequently, Nodes 2 and 3 retain the monotonically increasing IRT model in the ERSUTree model, echoing the logic applied in the IRTree model (Jin et al., 2022).

Combining the aforementioned details, when the four-point Likert scale is transformed into three-node data (as shown in Figure 1), the response probabilities at each node are as follows:

In Node 1 ($k = 1$), the probability of person j choosing to agree with item i is calculated as:

$$P(Y_{ijk} = 1|\theta_j) = \frac{\psi(\tau_{ik})}{\psi\left[\alpha_{ik}(\theta_j - \delta_{ik})\right] + \psi(\tau_{ik})} \quad (11)$$

For the individual j responding to item i, the probability of choosing extreme expression items such as "strongly disagree" or "strongly agree" is given by:

$$P(Y_{ijk} = 1|\eta_j) = \frac{\exp[\alpha_{ik}(\eta_j - \beta_{ik})]}{1 + \exp[\alpha_{ik}(\eta_j - \beta_{ik})]}. \quad (12)$$

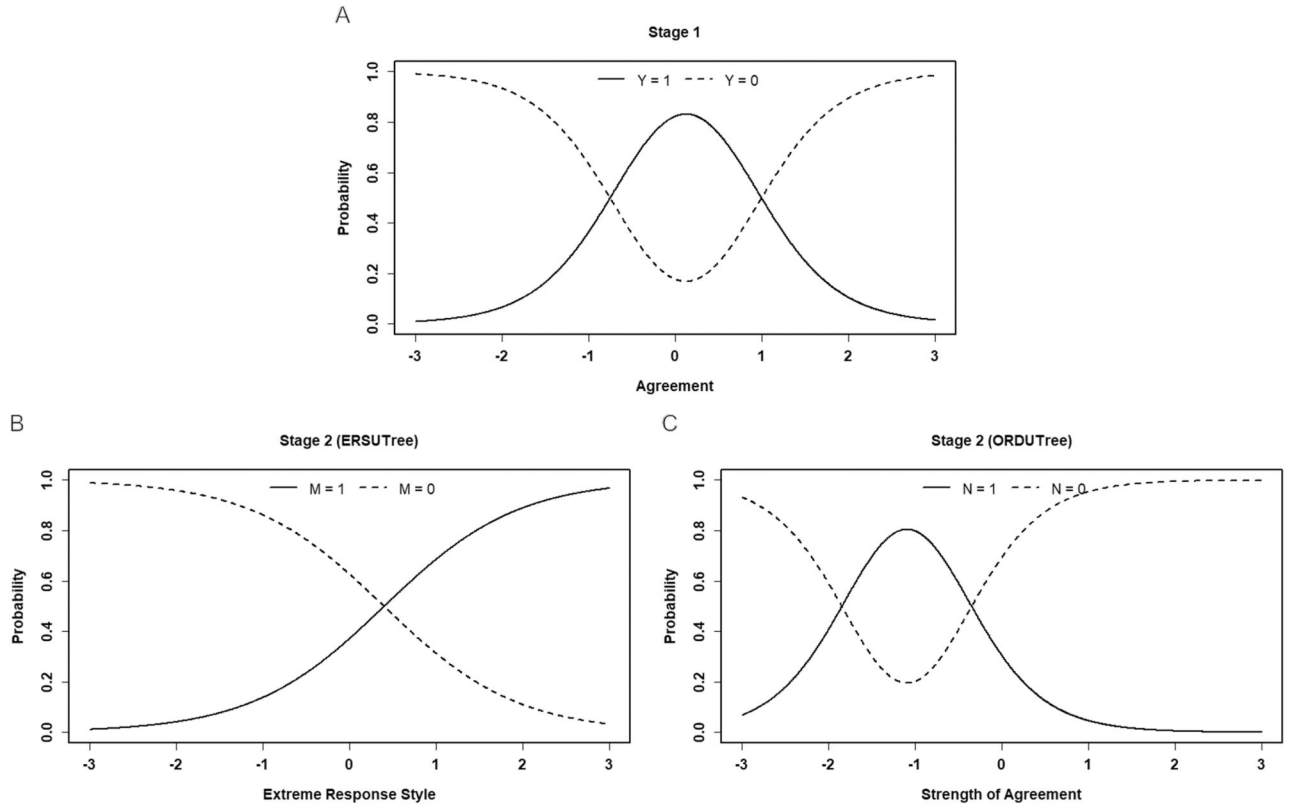

Figure 4. Diagram illustrating response probabilities across decision stages for the ERSUTree and ORDUTree models. *Note*. In Stage 1, all UTree models follow the same ideal point process where $Y = 1$ denotes a tendency to agree with the item statement, and $Y = 0$ indicates a tendency to disagree. In Stage 2, the ERSUTree model hypothesizes decisions based on ERS: $M = 1$ signifies extreme expression options ("strongly agree" or "strongly disagree"), while $M = 0$ represents non-extreme expressions ("somewhat agree" or "somewhat disagree"). The ORDUTree model hypothesizes decisions based on the target trait: $N = 1$ denotes options with a stronger intensity of agreement ("somewhat disagree" or "strongly agree"), and $N = 0$ indicates options with a lower intensity of agreement ("strongly disagree" or "somewhat agree").

In general, the probability of individual j to choose a specific option in item i is as follows:

$$\begin{aligned} P(Y_{ij} = 1|\theta_j, \eta_j) &= P\left(Y_{ij1} = 0|\theta_j\right) \times P\left(Y_{ij2} = 1|\eta_j\right) \\ P(Y_{ij} = 2|\theta_j, \eta_j) &= P\left(Y_{ij1} = 0|\theta_j\right) \times P\left(Y_{ij2} = 0|\eta_j\right) \\ P(Y_{ij} = 3|\theta_j, \eta_j) &= P\left(Y_{ij1} = 1|\theta_j\right) \times P\left(Y_{ij3} = 0|\eta_j\right) \\ P(Y_{ij} = 4|\theta_j, \eta_j) &= P\left(Y_{ij1} = 1|\theta_j\right) \times P\left(Y_{ij3} = 1|\eta_j\right) \end{aligned} \tag{13}$$

Ordinal unfolding tree model

The ORDUTree model, akin to the ORD model, follows an ordinal process but is distinctively built on the ideal point response process across two stages, leading to significant differences in its foundational concepts and response mechanisms. This model is divided into two distinct types. The first, ORDUTree.1, assumes respondents base all decision stages on a singular target trait. The second, ORDUTree.2, posits that respondents possess two different target traits during different decision stages. In the following sections, we will delve into both of these model types in detail.

ORDUTree.1 model

This model hypothesizes that there is one latent trait, θ, influencing decision-making across multiple stages. This is similar to the ORD.1 model, but in this model, respondents adhere to the ideal point cognitive decision-making process at all three nodes, which is fundamentally different from the response process assumed by the ORD IRTree. In this multi-process model, individuals make decisions across two stages, where respondents decide between agreeing or disagreeing with the item statement at Node 1, and choose between two options that represent a higher degree of agreement ("somewhat disagree" or "strongly agree") and a lower degree of agreement ("strongly disagree" or "somewhat agree") at Nodes 2 and 3, based on the same target traits. This means that the individual's choices on whether to agree with the item and the degree of agreement are highly

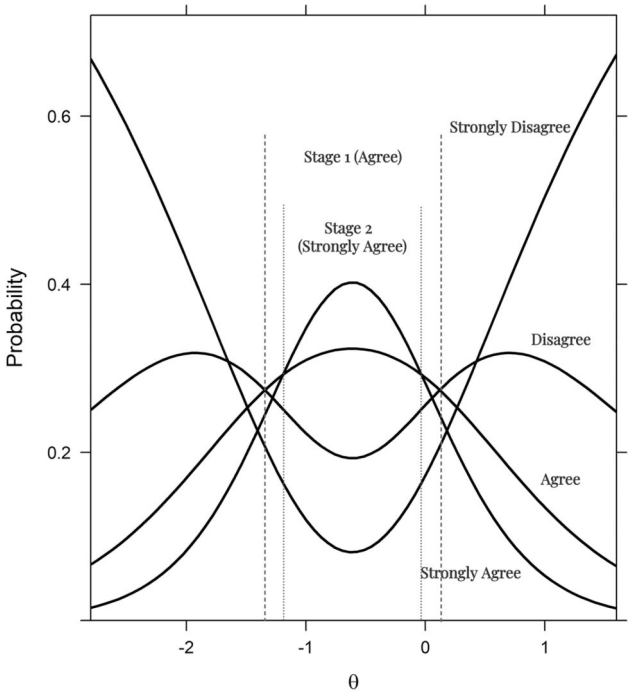

Figure 5. Diagram explaining the two-stage response process of a 4-point Likert scale with Ordinal Unfolding Tree model. *Note*. The Ordinal Unfolding Tree Model assumes that individuals undergo a two-stage decision-making process. In Stage 1, individuals decide whether to agree with the item statement, with the vertical range in Stage 1 representing a tendency to choose the 'agree' option. In Stage 2, the decision is about the degree of agreement, where the vertical range in Stage 2 indicates a tendency to choose the stronger agreement option of 'strongly agree'.

consistent, such as showing a high degree of agreement at Node 1 and similarly strong agreement at Node 3, indicating a strong preference for the "strongly agree" option; similarly, a strong preference for "strongly disagree" might show a similar pattern. This shows that when the target traits influencing choices in these two stages are highly correlated or consistent, using one target trait (θ) can adequately represent this similar two-stage decision process, which is also quite common and entirely possible in practice.

However, particularly, since this model involves only one target trait following the ideal point process across different decision stages, when the trait level is very close to the item location, the individual tends to agree with the statement at Node 1 and shows a very high degree of agreement at Node 3, eventually opting for the "strongly agree" option; conversely, when the trait level is very far from the item location, the individual shows a very low degree of agreement at Nodes 1 and 2, tending to choose the "strongly disagree" option. Combining the above, we find that this trait level aligns with our traditional understanding of the unfolding model. Therefore, we define this target trait similarly to the "attitudes towards the statement" in the unfolding model, where a smaller distance between the trait θ and the item location indicates a higher degree of agreement with the statement (strongly agree), and a larger distance indicates a lower degree of agreement (strongly disagree). The probabilities of respondents' choices across both decision-making stages are depicted specifically in Figure 4A and 4C. The slight distinction from the illustrations is that this model assumes respondents base their decisions at all stages on the same target trait.

In the ORDUTree.1 model, as per the ORD model, the four-point Likert scale can be transformed into three-node pseudo item data (see Figure 2). After transformation, across all nodes ($k = 1$, 2, and 3), the probability for individual j to choose an option representing a higher degree of agreement when responding to item i is as follows:

$$P\left(Y_{ijk}=1|\theta_j\right)=\frac{\psi(\tau_{ik})}{\psi\left[\alpha_{ik}\left(\theta_j-\delta_{ik}\right)\right]+\psi(\tau_{ik})} \tag{14}$$

Ultimately, the probability for individual j to choose a specific option when responding to item i is as follows:

$$\begin{aligned}
P(Y_{ij}=1|\theta_j)&=P\left(Y_{ij1}=0|\theta_j\right)\times P\left(Y_{ij2}=0|\theta_j\right)\\
P(Y_{ij}=2|\theta_j)&=\left(Y_{ij1}=0|\theta_j\right)\times P\left(Y_{ij2}=1|\theta_j\right)\\
P(Y_{ij}=3|\theta_j)&=P\left(Y_{ij1}=1|\theta_j\right)\times P\left(Y_{ij3}=0|\theta_j\right)\\
P(Y_{ij}=4|\theta_j)&=P\left(Y_{ij1}=1|\theta_j\right)\times P\left(Y_{ij3}=1|\theta_j\right)
\end{aligned} \tag{15}$$

ORDUTree.2 model

While the ORDUTree.1 model is suitable for individuals who make decisions based on highly similar or consistent target traits across stages, it becomes inappropriate when the individual's choices in the two stages are based on inconsistent target traits. For instance, if an individual tends to agree with the statement at Node 1, indicating a close distance to the item location with θ_1, but opts for a lower degree of agreement, such as "somewhat agree" at Node 3, which reflects a greater distance from the item location with θ_2, this discrepancy indicates a clear inconsistency between the two target traits, rendering the ORDUTree.1 assumption of a single trait inappropriate. Similarly, if an individual tends to disagree with the statement at Node 1 but opts for a higher degree of disagreement, like "somewhat disagree," at Node 2, this also demonstrates a significant difference in the latent traits at the two nodes.

Consequently, we introduce the ORDUTree.2 model, which allows respondents to base their decisions on different target traits at different stages. This model effectively accommodates the complexity of unfolding decision processes by separately estimating the different latent traits that influence various aspects of response behavior. In this model, individuals decide whether to agree or disagree with the item statement at Node 1, determining the overall direction of agreement or disagreement. Thus, we define the first stage's trait, θ_1, as the "direction of agreement". When θ_1 is close to the item location, the individual tends to agree with the statement; conversely, when it is distant, the tendency is to disagree. At Nodes 2 and 3, after determining the direction of agreement, the individual selects options that indicate either a higher or lower degree of agreement. At this stage, when θ_2 is close to the item location, the individual expresses a higher degree of agreement, choosing options indicative of strong agreement (such as "strongly agree" or "somewhat disagree"). Therefore, defining the target trait θ_2 as the "degree of agreement" at this stage is appropriate. The closer θ_2 is to the item location, the higher the degree of agreement; conversely, the farther it is, the lower the degree of agreement. These two distinct target traits might be related, or they could be unrelated. The decision-making process and the probability calculations by which respondents choose a specific option in this model are identical to those in the ORDUTree.1 model; however, the target trait upon which decisions are based varies by stage, as specifically depicted in Figures 4A & 4C.

The ORDUTree.2 model uses the same method of transforming a four-point Likert item to pseudo-items as the ORDUTree.1 model.

In Node 1 (k=1), the probability that individual j chooses to agree when responding to item i is given by:

$$P\left(Y_{ijk}=1|\theta_{j1}\right)=\frac{\psi(\tau_{ik})}{\psi\left[\alpha_{ik}\left(\theta_{j1}-\delta_{ik}\right)\right]+\psi(\tau_{ik})} \tag{16}$$

In Nodes 2 and 3 (k = 2 or 3), the probability that individual j responds to item i by choosing the option representing a higher degree of agreement ("somewhat disagree" or "somewhat agree") is given by:

$$P\left(Y_{ijk}=1|\theta_{j2}\right)=\frac{\psi(\tau_{ik})}{\psi\left[\alpha_{ik}\left(\theta_{j2}-\delta_{ik}\right)\right]+\psi(\tau_{ik})} \tag{17}$$

Meanwhile, θ_{j1} represents the direction of agreement, while θ_{j2} represents the degree of agreement. Respondents are more likely to choose the option indicating a higher degree of agreement when their position parameters θ_{j1} and θ_{j2} are closer to the item's position parameters δ.

Under the ORDUTree.2 model, the probability of respondent j selecting option m for item i is as follows:

$$\begin{aligned}
P(Y_{ij}=1|\theta_{j1},\theta_{j2})&=P\left(Y_{ij1}=0|\theta_{j1}\right)\times P\left(Y_{ij2}=0|\theta_{j2}\right)\\
P(Y_{ij}=2|\theta_{j1},\theta_{j2})&=P\left(Y_{ij1}=0|\theta_{j1}\right)\times P\left(Y_{ij2}=1|\theta_{j2}\right)\\
P(Y_{ij}=3|\theta_{j1},\theta_{j2})&=P\left(Y_{ij1}=1|\theta_{j1}\right)\times P\left(Y_{ij3}=0|\theta_{j2}\right)\\
P(Y_{ij}=4|\theta_{j1},\theta_{j2})&=P\left(Y_{ij1}=1|\theta_{j1}\right)\times P\left(Y_{ij3}=1|\theta_{j2}\right)
\end{aligned} \tag{18}$$

This probability function is akin to that in the ORDUTree.1 model. The primary difference lies in the basis of decisions at different nodes: decisions at Node 1 are predicated on the trait θ_{j1}, while decisions at Nodes 2 and 3 rely on θ_{j2}.

The ORDUTree model, akin to the ORD model, adheres to an ordinal process but is distinctively built on the ideal point response process across two stages, introducing significant differences in its foundational concepts and response mechanisms. The ORD models assume that a larger value of θ inclines individuals toward options with a higher degree of agreement. In contrast, the ORDUTree model follows the ideal point response process, where the smaller the distance between θ and the item location, the higher the degree of agreement selected, reflecting significant differences in underlying hypotheses and latent response process.

Simulation research

To evaluate the performance of the UTree models in comparison to IRTree models, this section intends to generate response data under various conditions through simulations. The generated data are then analyzed across a total of six models. Subsequently, the performances of these models are assessed with established metrics. The primary research questions to be addressed include: Firstly, can the magnitude of fit indices accurately determine which measurement model aligns best with the response data? Secondly, when the estimation model mirrors the true model, can the IRTree and UTree models accurately recover the parameters? And if a mis-specified model is applied, what repercussions ensue from utilizing such a misaligned model? Finally, how do the IRTree and UTree models fare under varied conditions, and which factors might potentially impact their performance?

Data generation

To assess the performance of UTree models under various conditions, including those of IRTree and UTree, and to more accurately gauge the properties and stability of these models, as well as to determine if fit indices can accurately discern the models that better fit the response data, we generated simulated data based on all six models—ERS, ORD.1, ORD.2, ERSUTree, ORDUTree.1, and ORDUTree.2—under a range of conditions.

It is worth further explaining that the primary objective of this study is to discuss the differences between UTree and nested IRTree models. Although many studies have found that the nested IRTree models outperform traditional single-decision GRM and linear IRTree models in Likert scales (Böckenholt & Meiser, 2017; Jeon & De Boeck, 2016; LaHuis et al., 2019; Tijmstra et al., 2018). Among the few existing IRTree models considering the ideal point decision process, this model also outperforms single-decision GGUM and GRM models (Jin et al., 2022). Considering the above and the text length, GRM, GGUM, and linear IRTree models were not included in the comparisons within the simulation and empirical studies of this paper. These models can be further explored in future research.

Simulation conditions

This study uses a four-point Likert scale as its foundation. For the four models that contain two latent traits, namely ERS, ORD.2, ERSUTree, and ORDUTree.2, we simulate response data under a total of 27 different conditions for each model, broken down as 3 (sample size) × 3 (test length) × 3 (latent trait correlation). For the ORD.1 and ORDUTree.1 models, which contain only one latent trait, there is no need to consider the factor of latent trait correlation. Thus, for each of these models, response data was simulated under 9 different conditions, delineated as 3 (sample size) × 3 (test length). The detailed simulation conditions are as follows:

Sample size (N). Three levels are considered: small (500 respondents), medium (1,000 respondents), and large (2,000 respondents). These conditions align with existing unfolding model research (Roberts & Laughlin, 1996).

Test length (I). Three lengths are considered: short (5 items), medium (10 items), and long (20 items). Given that test length can significantly influence the estimation accuracy of unfolding models (Roberts & Laughlin, 1996), this study has set up scenarios with three different test lengths.

Latent trait correlation (ρ). Three degrees of correlation are explored: no correlation ($\rho_{\theta_1\theta_2(\eta)} = 0$), low correlation ($\rho_{\theta_1\theta_2(\eta)} = 0.3$), and high correlation ($\rho_{\theta_1\theta_2(\eta)} = 0.6$). Some previous research considered correlations between different traits at different nodes (Böckenholt, 2017; Böckenholt & Meiser, 2017), while other studies operated under the assumption of no correlation (Kim & Bolt, 2021). Thus, this study considers various correlation extents to assess model efficacy.

Subsequently, based on this foundation, we determine the distribution of each parameter for the simulated data.

Discrimination parameter (α). Parameters are drawn from a uniform normal distribution in the range [0.5, 2.0], as outlined in the study by Kim & Bolt (2021).

Item location parameter (δ or β). Parameters are generated from a truncated normal distribution in the range [−2, 2], which is similar to the parameter ranges suggested in previous studies (Andrich, 1988; Kim & Bolt, 2021; Roberts & Laughlin, 1996).

Threshold parameter (τ). Parameters for threshold values are selected from a uniform distribution between [−2, −0.5], mirroring the findings from unfolding models in previous research (Roberts & Laughlin, 1996).

Person location (θ) and extreme response style parameter (η). Parameters are generated from a normal distribution with a mean of 0 and a standard deviation of 1, consistent with prior research (Andrich, 1988; Kim & Bolt, 2021; Roberts & Laughlin, 1996).

Data generation process

After determining the simulation conditions and the distribution of each parameter, simulated datasets are generated through the following steps:

Step 1: Following the pre-determined sample size, test length, and latent trait correlation conditions, respondents are generated under each condition based on the six different models.

Step 2: Following the established parameter conditions, the three-node discrimination parameter, item location parameter, threshold parameter, person location parameter, and response style parameter are generated.

Step 3: Based on the six different types of IRTree and UTree models, apply the parameters and conditions generated in Step 1 and Step 2 to the specific models. Compute the probability for each respondent to choose different category options m ($m = 1, 2, 3, 4$).

The formula for probability calculation can be found in the model settings section.

Step 4: Based on the option probabilities from the previous step, generate multinomial responses.

Step 5: Convert the simulated response data into three-node pseudo-items based on the analysis model settings. Use the conversion methods outlined in Table 1 for ERS and ERSUTree, and Table 2 for ORD and ORDUTree models.

Step 6: Repeat Steps 4 and 5, generating 100 datasets under each condition for replication analysis purposes.

Analysis procedure

We analyze the simulated datasets using the standard EM algorithm with fixed quadrature *via* the *mirt 1.38.1* package in R (Chalmers, 2012). Following the approach of Li et al. (2025), we wrote the data simulation and analysis code for all six IRTree and UTree models, which is hosted on the Open Science Framework (available at: https://osf.io/t8znm/). Model fit metrics, item parameter recovery, and the proportion of true values contained within the 95% confidence intervals for θ and η estimated parameters are chosen as evaluation metrics. The chosen method of estimation significantly reduces computation time. Even for the most complex UTree models, convergence is usually achieved within a few to tens of minutes. This greatly enhances computational efficiency, increasing the practicality and operability of both the IRTree and UTree models. In contrast, the Bayesian Markov Chain Monte Carlo method, used by other studies (Jin et al., 2022; Kim & Bolt, 2021), typically requires several hours for estimation, e.g., 9 h for the complex UTree models.

In the model evaluation metrics, fit metrics primarily consist of the Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978) and the Sample-Size Adjusted Bayesian Information Criterion (SABIC) (Sclove, 1987). It is worth noting that compared to the commonly used BIC, the SABIC places a penalty for adding parameters based on sample size, and previous research has demonstrated that SABIC is particularly suitable for model comparison and should be prioritized in such contexts (Chen et al., 2017; Enders & Tofighi, 2008; Jeon & De Boeck, 2019a; Jeon & De Boeck, 2019b). Therefore, we will utilize this metric in our study.

The item parameter recovery is mainly used to assess the accuracy and precision of the parameters estimated by the model, including both Bias and RMSE values. The accuracy of individual parameter estimates, specifically for θ and η, is gauged by the proportion of times the true values are encompassed within the 95% confidence interval of these estimates (Kim & Bolt, 2021). Essentially, a higher proportion indicates that the estimated confidence intervals are consistently capturing the true values, denoting a reliable estimation process. Additionally, the Bias and RMSE of these person parameter estimates will also be presented to describe the precision of these models.

It is noteworthy that throughout the estimation process of the UTree model, we employ the GGUM model at each stage. This model exhibits the characteristic of reflective invariance. This means that the likelihood of a set of responses, given θ and δ vectors, is identical to the likelihood given vectors $-\theta$ and $-\delta$ (Bafumi et al., 2005). Therefore, when employing Bayesian estimation for unfolding models, estimating with one set of response data can potentially lead to two scenarios: one with θ and δ and another with $-\theta$ and $-\delta$, which lead to non-convergent estimation results and identification issues (Duck-Mayr & Montgomery, 2023). Such challenges severely complicate the estimation and application of the GGUM model. Yet, the existing research on unfolding models rarely addresses this issue (de la Torre et al., 2006; Jin et al., 2022; Roberts et al., 2000; Roberts & Laughlin, 1996).

The *mirt* package utilizes the standard EM algorithm and achieves stable and convergent results when analyzing the GGUM model (Chalmers, 2012). However, there can still be instances where our simulated data uses θ and δ, but the estimation may result in $-\theta$ and $-\delta$. To obtain reliable results regarding the accuracy of the estimated positions for both individuals and items, we drew upon existing research addressing reflective invariability (Duck-Mayr & Montgomery, 2023; Stephens, 1997). Our approach involves first calculating the overall correlation between all estimated individual parameters θ (η) and all item parameters δ with their respective true values. Then, based on the sign of the correlation, we adjust the estimated values to align with the direction of the true values (by multiplying them by either 1 or -1). Finally, we compute the Bias and RMSE values, enabling us to assess the genuine accuracy of our estimations.

Results

Fit indices

Initially, based on the data generated from six distinct models, Table 3 presents the average values of AIC and SABIC for each model across all conditions over

Table 3. Average fit index values (AIC and SABIC) and the percentage of times each model exhibits the lowest fit index values across all conditions for six models.

		Data generation model					
Fitted model	Index	ERS	ORD.1	ORD.2	ERS UTree	ORD UTree.1	ORD UTree.2
ERS	AIC	**30697.14 (99%)**	20677.56 (0%)	31718.33 (0%)	31514.02 (2%)	34320.08 (1%)	34606.69 (0%)
	SABIC	**30820.11 (100%)**	20763.84 (0%)	31841.31 (0%)	31637.00 (4%)	34443.05 (1%)	34729.66 (0%)
ORD.1	AIC	32203.81 (0%)	**20042.20 (100%)**	31635.71 (0%)	32757.35 (0%)	34110.49 (4%)	34686.63 (0%)
	SABIC	32325.05 (0%)	**20126.85 (100%)**	31756.95 (0%)	32878.59 (0%)	34231.73 (4%)	34807.87 (0%)
ORD.2	AIC	31743.72 (0%)	–	**30675.23 (100%)**	32437.13 (0%)	34096.46 (1%)	34416.04 (0%)
	SABIC	31866.69 (0%)	–	**30798.20 (100%)**	32560.11 (0%)	34219.43 (1%)	34539.02 (2%)
ERSUTree	AIC	30741.74 (1%)	–	31765.79 (0%)	**30991.75 (98%)**	–	34108.68 (0%)
	SABIC	30884.92 (0%)	–	31908.97 (0%)	**31134.93 (96%)**	–	34251.86 (0%)
ORDUTree.1	AIC	31189.48 (0%)	20119.18 (0%)	31001.46 (0%)	31637.21 (0%)	**32688.64 (95%)**	33993.89 (0%)
	SABIC	31371.35 (0%)	20246.14 (0%)	31183.33 (0%)	31819.08 (0%)	**32870.51 (94%)**	34175.75 (0%)
ORDUTree.2	AIC	30803.08 (0%)	–	30814.21 (0%)	31063.20 (0%)	–	**33376.22 (99%)**
	SABIC	30986.67 (0%)	–	30997.80 (0%)	31246.80 (0%)	–	**33559.81 (98%)**

Note. Each condition was replicated 100 times (consistent with the subsequent simulation analysis).

100 replications. Furthermore, the table showcases the proportion of instances where the AIC and SABIC values for a particular model are the lowest among the six models, indicating the best model fit. Statistical power denotes the frequency with which the true model is accurately identified. A value approaching 1 (or 100%) signifies superior statistical efficacy. The data in Table 3 demonstrates that regardless of whether the data originates from the IRTree or UTree models, both AIC and SABIC consistently exhibit high power across varied item numbers, sample sizes, and trait correlations. They consistently return the smallest estimated values, correctly identifying the genuine data model. Even for the data generated from the more intricate UTree model, the AIC and SABIC maintain a power close to 1, accurately discerning the correct model that aligns with the respondent's decision-making process. This underscores the viability of AIC and SABIC in pinpointing the true model underlying response data. We also presented the BIC estimates in Appendix A (Table A1). While this metric performed well in estimating IRTree models, it showed much poorer performance in estimating the more complex UTree models, especially for the ERSUTree and ORDUTree.2 models. Even when the correct model was used, the misclassification rate was as high as 11%. Therefore, we chose not to use this metric as a model selection criterion in the subsequent analysis.

Interestingly, whether the data is generated from IRTree or UTree model based on a single trait, UTree models that assume two different latent traits (ERSUTree and ORDUTree.2) exhibit excellent sensitivity, which refers to their capacity to correctly identify when response data does not conform to the measurement assumptions of the model. We found that these two model types often suffer from computation issues under the single trait data generation condition, reporting estimation errors and convergence problems. This observation indicates there is no need for further examination of AIC and SABIC values for these models and directly indicates that respondents' answers are not based on two latent traits under conditions where data is generated from a single trait. Notably, UTree models that assume two traits struggle due to a lack of information in the second dimension to estimate parameters beyond those defined by the data generation model (which involves a single latent trait). As a result, the estimation algorithm becomes stuck on a likelihood plateau, making it extremely difficult, if not nearly impossible, to achieve accurate estimates and avoid convergence problems. Additionally, from a more rigorous perspective, the validity of this indicator in Bayesian estimation needs further examination. Therefore, sensitivity can be used as an auxiliary judgment tool, while fit indices might still be the more crucial indicators for determining whether the model matches the data.

In contrast, the sensitivity of IRTree models that assume two latent traits falls short. The ERS model cannot directly indicate during its estimation whether the respondents are answering based on a single latent trait through the estimation process. Instead, further evaluation and comparison of AIC and SABIC values with other models are required. The ORD.2 model can encounter estimation errors and convergence issues and such incidences indicate the respondents' answers may not be based on two latent traits only when respondents are based on a single trait and follow a dominance process (i.e., the ORD.1 model). However, when respondents answer based on a single trait but follow an ideal point process (i.e., the ORDUTree.1 model), this model cannot directly diagnose whether the respondents' answers are based on a single latent trait. This necessitates additional scrutiny of AIC and SABIC and comparison with other models to make such a determination, such as Table 3, rendering the process less efficient.

Item parameter recovery

Based on data generated from six distinct models, we compute the average Bias and RMSE estimation results for all models. It is important to note that because the six models have different model settings at various decision stages, the item parameters differ in meaning and are therefore not directly comparable. Consequently, we only present item parameter recovery results where the data generation model matches the estimation model. The estimation results for the six models with consistent data generation and estimation can be found in Figures 6 and 7. To ensure that the graphics are both concise and readable, we display two figures showcasing the average Bias and RMSE values for the item parameters in each model under various conditions of test length, sample size, and latent trait correlations. Specific numerical details are provided in Appendix B (Tables B1–B5). For more granular Bias and RMSE estimation results for each item parameter within each model, please refer to Appendix C (Tables C1).

Overall, combining the aforementioned results, we observe that when the estimation model matches the data generation model (i.e., the correct model), all IRTree and UTree models can accurately retrieve item parameters. The average Bias value for each item parameter approaches zero, and RMSE values typically range between 0.1 and 0.2. The distribution of different item parameters is also relatively concentrated, with no item parameters exhibiting particularly high Bias and RMSE. Even for the UTree model which is complex in its formulation, its parameter recovery remains commendable. The Bias for its parameters is close to zero, and RMSE is stable around 0.18, suggesting satisfactory recovery of data-generating values of the model parameters.

Next, we delve deeper into the influence of different conditions on the performance of IRTree and UTree models. Under varying test length conditions, with increasing numbers of items, almost all six models show a significant downward trend in RMSE, and their distributions become more concentrated. This indicates that the absolute deviations in the estimation of different items are generally reduced, leading to more accurate estimation results. Regarding Bias, the average Bias in IRTree models remains relatively unchanged, but the distribution of Bias values across different items becomes wider, especially when the number of items reaches 20. It is important to note that the ORD.1 model encounters computational issues and reports convergence problems when the number of items is 20 and the sample sizes are 1000 or 2000 (as detailed in Appendix Table B2). Consequently, only results for the sample size of 500 are presented under the 20-item condition. Although the Bias and RMSE values in this condition remain relatively high, they still exhibit a downward trend compared to the conditions with 5 or 10 items, aligning with the overall patterns observed in other models. As for UTree models, the mean and distribution of Bias are stable, indicating that the relative bias of these models is less affected by changes in the number of items and remains consistent. Although UTree models may seem significantly more complex than IRTree models, their stability and estimation accuracy are particularly noteworthy. Overall, with 10 items, most IRTree and UTree models can achieve smaller Bias and RMSE results, demonstrating their effectiveness across a range of configurations.

Across different sample sizes, as the sample size increases, all IRTree and UTree models show a significant downward trend in both Bias and RMSE values, with their distributions becoming more concentrated, indicating a general reduction in overall item deviations. When the sample size reaches 1000, the RMSE for IRTree models drops to around 0.15, while the more complex UTree models maintain an excellent RMSE of about 0.18. The mean Bias values for all these models hover around zero, with minimal differences between items, indicating that these models can provide quite accurate estimates.

In scenarios with varying latent trait correlations, both IRTree and UTree models show little change in

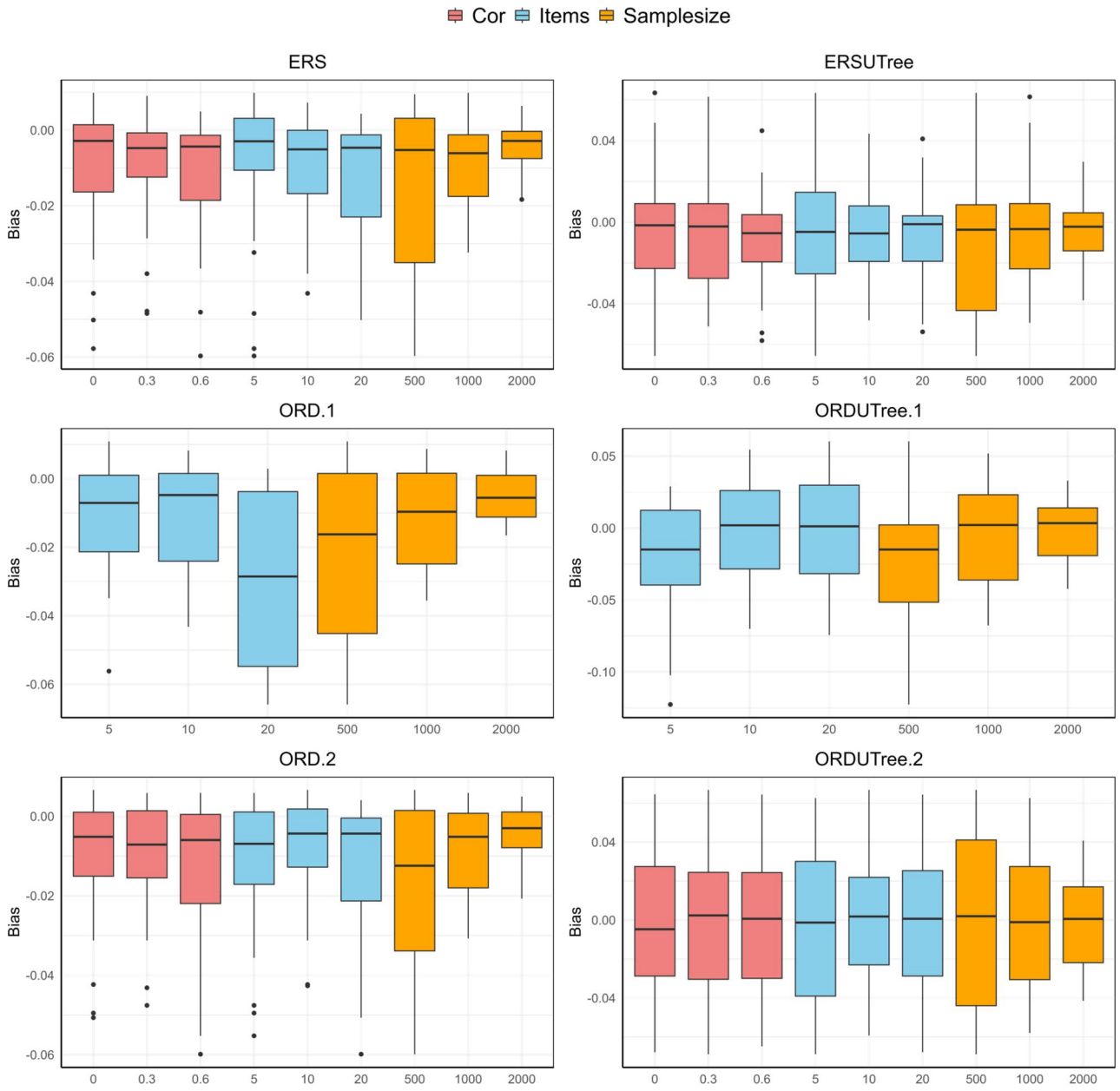

Figure 6. Average Bias of estimated item parameters across all simulated conditions when the estimated model and data generation model are consistent.

the mean and distribution of RMSE and Bias. The RMSE means stay around 0.15 for IRTree and 0.18 for UTree, with Bias concentrated near zero. This indicates that these models, which can freely estimate different traits at different stages, can effectively handle parameter estimation under various trait correlation conditions, returning accurate item parameters.

Proportion of 95% confidence intervals of estimated θ and η parameters containing true values

Table 4 presents the average proportion, across all simulated conditions, of the 95% confidence intervals for θ and η parameter estimates that encompass the true values. When the estimation model is the correct one, all six models can accurately retrieve respondents' target traits or response style parameters, with the proportion of 95% confidence intervals containing the true values being exceptionally high. Additionally, it is worth noting that the results in Table 4 indicate that the standard errors (SEs) obtained from the correctly specified model estimates are relatively low across all six models. However, the fact that the 95% confidence intervals formed from these estimates encompass the true values at the highest proportion suggests that this is not due to large estimated standard errors. Rather, it indicates that the target traits

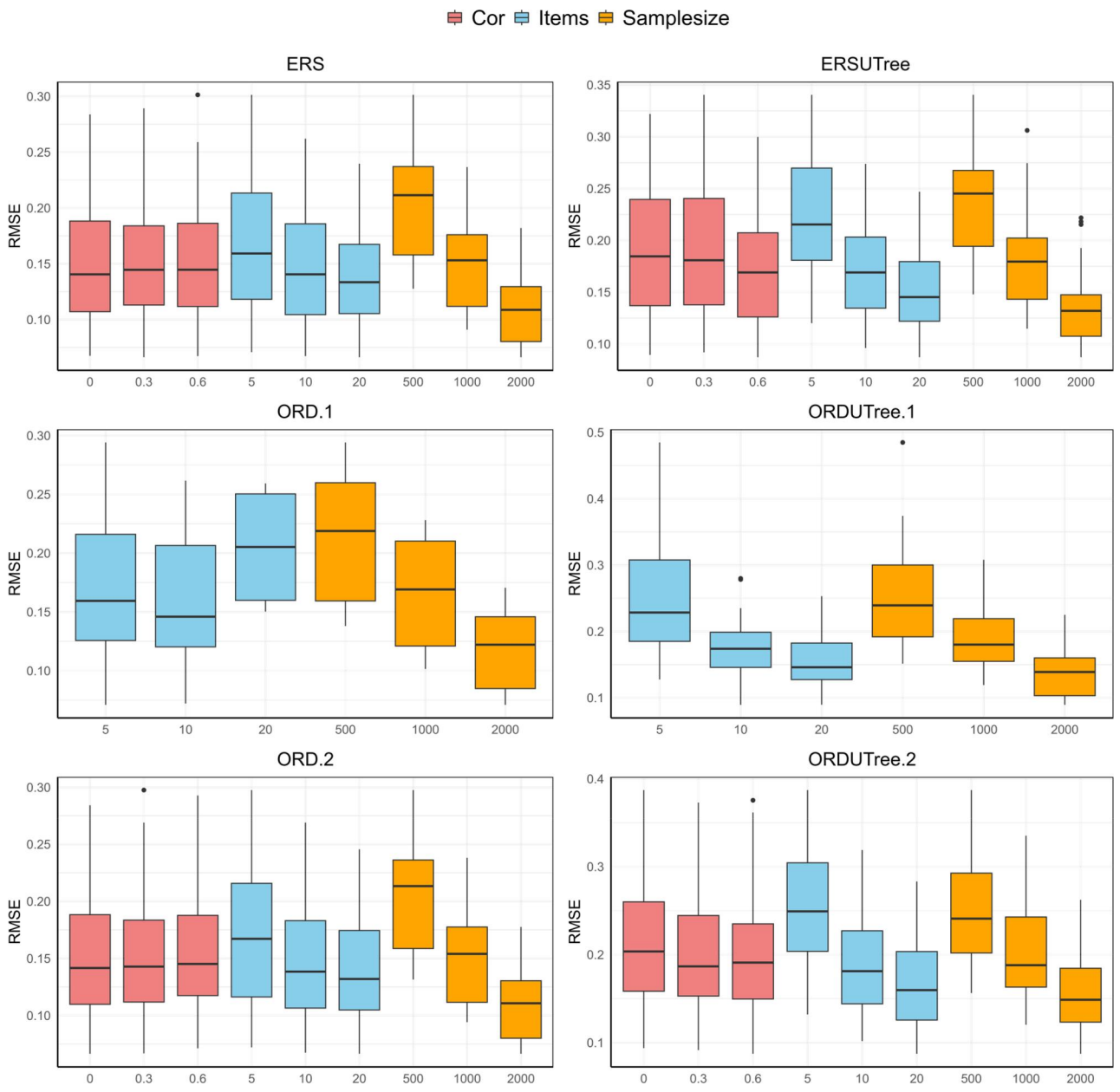

Figure 7. Average RMSE of estimated item parameters across all simulated conditions when the estimated model and data generation model are consistent.

Table 4. Average percentage of 95% confidence interval for estimated θ and η includes the true value across all conditions and replications.

	Fitted model									
	ERS		ORD.1	ORD.2		ERS UTree		ORD UTree.1	ORD UTree.2	
Data generation model	θ	η	θ	θ_1	θ_2	θ	η	θ	θ_1	θ_2
ERS	**0.95**	**0.95**	0.91	0.95	0.92	0.89	0.95	0.58	0.88	0.92
	(0.53)	**(0.53)**	(0.5)	(0.53)	(0.64)	(0.6)	(0.53)	(0.44)	(0.6)	(0.52)
ORD.1	0.94	0.95	**0.95**	–	–	–	–	0.85	–	–
	(0.51)	(0.54)	**(0.46)**					(0.55)		
ORD.2	0.95	0.92	0.79	**0.95**	**0.95**	0.86	0.91	0.60	0.82	0.82
	(0.53)	(0.64)	(0.48)	**(0.53)**	**(0.53)**	(0.62)	(0.64)	(0.52)	(0.69)	(0.67)
ERSUTree	0.89	0.95	0.79	0.88	0.93	**0.94**	**0.95**	0.70	0.94	0.92
	(0.68)	(0.56)	(0.61)	(0.68)	(0.66)	**(0.56)**	**(0.55)**	(0.48)	(0.56)	(0.52)
ORDUTree.1	0.88	0.89	0.82	0.84	0.84	–	–	**0.94**	–	–
	(0.62)	(0.65)	(0.58)	(0.61)	(0.62)			**(0.45)**		
ORDUTree.2	0.88	0.90	0.79	0.88	0.86	0.94	0.90	0.80	**0.94**	**0.94**
	(0.68)	(0.73)	(0.63)	(0.69)	(0.69)	(0.57)	(0.72)	(0.56)	**(0.57)**	**(0.56)**

Note. The value in parentheses next to each θ and η represents the standard error (SE) of the latent traits obtained from the MCMC estimation.

estimated by these models are relatively accurate. Furthermore, when we delve deeper into different numbers of items, sample sizes, and latent trait correlation conditions, the proportion of 95% confidence intervals from the correct model containing the true values remains stable between 0.94 and 0.95. They effectively capture respondents' target traits or extreme response style parameters under almost all conditions, even when the number of items is very few ($I = 5$) or the sample size is quite low ($N = 500$).

However, when an incorrect estimation model is used, the proportion of the 95% confidence intervals for estimated θ and η parameters that cover the true values drops significantly. This drop is particularly evident when there is a misconception about the respondents' decision-making process. For instance, if respondents follow an "ideal point" response process but the "dominance" process-based IRTree model is erroneously used, the proportion of 95% confidence intervals containing the true personal parameter values is much lower. Additionally, if individuals make decisions based on two different types of target traits (θ_1 and θ_2) and mistakenly use models like ORD.1 or ORDUTree.1 that assume a single target trait, the proportion of 95% confidence intervals containing the true values for target trait θ also drops substantially. This indicates a significant bias in the estimation of respondents' latent abilities, making the results unreliable. These findings further attest to the importance of choosing the right model for estimating respondents' latent traits. Using an incorrect model can lead to biased estimation outcomes, greatly affecting researchers' accurate judgment of respondents' true latent traits.

Furthermore, to substantiate the validity of these model estimates in assessing target traits, we have included in Appendix D the recovery conditions for θ and η across all six models. The results in the table demonstrate that when the estimation model aligns with the data generation model, the RMSE values for the estimates of θ and η are the smallest. This is consistent with the highest proportions of 95% confidence intervals covering the true values, further affirming the reliability of using the correct model to estimate target traits.

Summary and discussion

Based on the above simulation results, we found that the AIC and SABIC indices demonstrate high statistical power in accurately identifying the correct model that represents the response process of response data. Furthermore, the ERSUTree and ORDUTree.2 models show a high level of sensitivity by promptly reporting estimation errors and convergence issues during the estimation process for respondents' data under a single trait condition. This direct detection capability allows us to ascertain that the responses are not based on two latent traits, thereby avoiding the additional time and resource expenditure that would be required for further AIC and SABIC comparisons.

When the correct model is utilized, all six categories of IRTree and UTree models adeptly retrieve item parameters. Additionally, as both the number of items and the sample size increase, there is a notable improvement in the precision of these models' estimations. These models also consistently maintain their stability across varying conditions of latent trait correlation. Moreover, we found that when the correct model is employed, both IRTree and UTree models are capable of accurately estimating respondents' latent traits, with the 95% confidence intervals for the θ and η parameter estimations substantially covering the true values, and exhibiting low Bias and RMSE values. Conversely, employing an incorrect model markedly reduces the accuracy of these intervals and introduces severely biased latent trait estimations for respondents. This issue is particularly pronounced when an erroneous response process model is used, such as applying IRTree models to data where respondents are answering based on an ideal point process.

Real data application

Through the analysis of responses from two empirical Likert-scale surveys, we aim to evaluate the performance of various IRTree and UTree models, and based on this, determine which model is best aligned with the empirical data. On this foundation, we seek to identify possible potential decision-making processes that respondents use in actual Likert-scale scenarios. Specifically, do they base their decisions on a dominance or an ideal point approach? When opting for extreme options in the second decision phase, is their choice driven by an extreme response style or specific target traits? And if it is the latter, do these target traits vary across different decision stages? With these insights in hand, we then delve deeper into item parameters and respondents' latent traits, utilizing the most fitting model to explore the intricacies of respondents' item response behaviors.

Example 1: reading interest

Datasets and analytical procedure

We utilize data from the "Reading Interest" section of the 2018 Programme for International Student Assessment (PISA) Middle School Student Questionnaire, which is publicly accessible *via* the official PISA website (OECD, 2018). This "Reading Interest" section comprises five items designed to gauge students' attitudes toward reading. Participants rated each item on a 4-point Likert scale: 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. Three of the items had a negative connotation (e.g., "For me, reading is a waste of time"), while the remaining two were phrased positively (e.g., "Reading is one of my favorite hobbies"). Students are likely to agree with items when their reading interest aligns closely with the item's statement. Previous research also found that employing the unfolding model for analyzing some of Likert scale data in PISA is more appropriate (National Center for Education Statistics, 2008). This model may better explain the data by capturing the ideal point response process and nuanced attitudes toward statements. For analytical consistency, we adjusted the scores of negatively framed items to align all item scores in the same direction.

In this study, we considered the responses from the students in Mainland China, resulting in a dataset of 11,832 student responses. We randomly selected 2,000 students out of 11,832 respondents for ease of estimation. Our earlier simulation study demonstrated that both the IRTree and UTree models perform well for a small number of items ($I = 5$), with sample sizes of $n = 1{,}000$ and 2,000.

All six models are applied to analyze the dataset using the "*mirt* 1.38.1" package in R. We employ relative fit indices (AIC and SABIC), absolute fit indices (h^2), item parameters, and latent traits to probe the effectiveness of the models and to delve into the students' response processes.

Results

Fit indices. Table 5 presents the relative fit indices for the six models. The UTree models display notably lower fit indices than the IRTree models, suggesting that the ideal point response process is more appropriate. Particularly, the ORDUTree.2 model boasts the smallest AIC and SABIC values, underscoring its prominence as the best-fitting model. Figure 8 presents the results of the absolute fit indices for six models. It's worth noting that common absolute fit indices such as M_2, S-X^2, RMSEA, and SRMR require that individuals have no missing response data (Chalmers, 2012). However, all IRTree and UTree

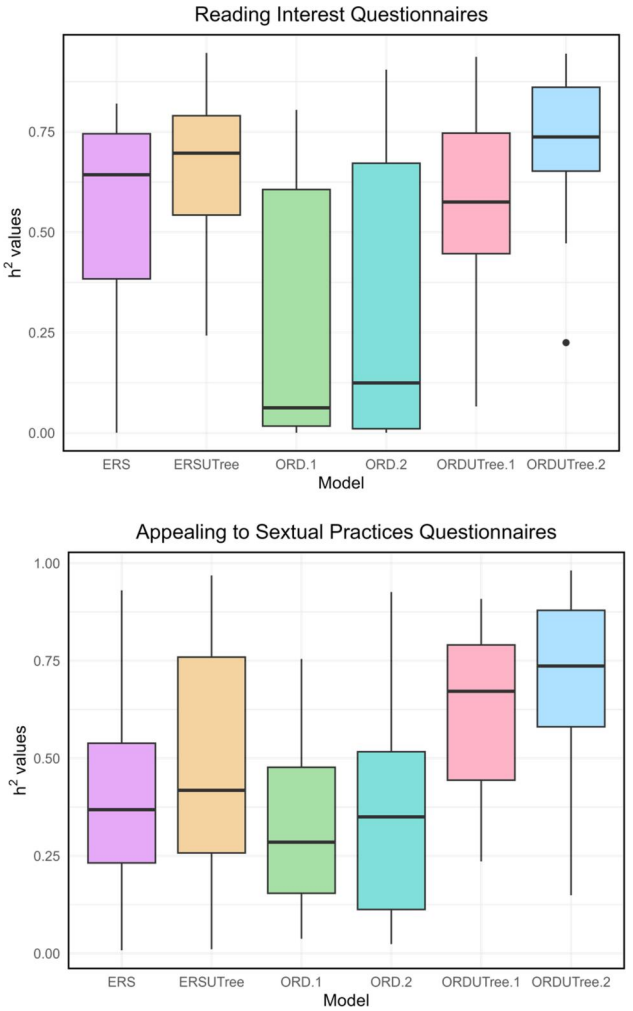

Figure 8. Absolute fit indices (h^2) for different IRTree and UTree models in "reading interest" and "appealing to sexual practices" questionnaires.

Table 5. Relative fit indices for the six estimated models in the "reading interest" and "appeal to sexual practices" questionnaires.

	ERS	ORD.1	ORD.2	ERS UTree	ORD UTree.1	ORD UTree.2
			Reading interest			
AIC	18420.45	19722.29	19582.05	17440.69	17634.52	17386.74
SABIC	18495.58	19795.00	19657.19	17527.95	17743.59	17498.24
			Appeal to sexual practices			
AIC	34101.32	34898.37	34116.97	33659.12	33834.53	32763.38
SABIC	34289.28	35084.26	34304.92	33878.06	34113.37	33044.28

model simulations utilize pseudo-item data where all individuals have missing data in either Node 2 or Node 3, making it impossible to calculate these indices. Therefore, we chose to report h^2 values (Factor Communality), where higher h^2 values represent greater explanatory power of the latent traits, which can more effectively explain the shared variance of the manifest items, indicating that the latent traits and the model construction are better able to explain the responses of the manifest items, thus suggesting a better fit. Figure 8 illustrates the fit distribution for all items across the six models through box plots, indicating that the ORDUTree.2 model also had the highest h^2 values across both datasets, with most items exhibiting high h^2 values. This suggests that the two factors of this model have the highest explanatory power, confirming that this model is the most congruent with the data in terms of data-factor fit.

This outcome indicates that when respondents select extreme options like "strongly agree" or "strongly disagree", their decisions are not primarily influenced by an extreme response style, but rather by target traits associated with reading interest. Additionally, respondents do not rely on a single reading interest-related target trait when responding to this questionnaire; they employ different target traits across decision-making phases. This finding suggests that for the "Reading Interest" questionnaire, student responses are influenced by two discrete latent traits associated with reading interest: a general agreement tendency in Node 1 and the degree of this agreement in Nodes 2 and 3.

Drawing from our simulation study's insights, if the data is premised on a sole latent target trait—be it through a dominant or ideal point response process—the ORDUTree.2 model suffers from serious estimation difficulties, often resulting in terminating computation. However, we were able to retrieve estimation results from this model. This indirectly reaffirms that, in this dataset, respondents do not hinge their responses on a singular target trait; instead, they base their answers on two distinct latent traits.

Item parameters and latent traits. With the ORDUTree.2 model proving the best fit, we delve into its estimated item parameters and student latent traits, as presented in Table 6. Most items display strong discrimination across all three nodes. In Node 1, the trait representing students' agreement with reading interest is moderate ($\theta_1 = -0.004$). This alignment between the student trait and the item and threshold parameters suggests a relatively high level of interest in reading. In Node 2, the degree of students' agreement ($\theta_2 = 0.559$) closely matched item and threshold parameters, indicating a preference for "disagree" over "strongly disagree" and signaling heightened reading interest. In contrast, Node 3 showcases a greater disparity between students' agreement intensity and the item and threshold parameters, leaning toward "agree" over "strongly agree". This indicates a relatively elevated—rather than the highest—level of reading interest. Overall, the model's outcomes highlight strong discrimination within the questionnaire, and students exhibited moderate to moderately high reading interest.

Table 6. Average estimated item parameters from ORDUTree.2 model for "reading interest" and "appeal to sexual practices" questionnaires.

	a_1	a_2	δ_1	δ_2	τ_1	τ_2	θ_1 (SE)	θ_2 (SE)
Reading interest								
Node 1	4.07		−0.31		1.06		−0.004 (−0.002)	
Node 2		2.83		−1.36		1.71		0.559
Node 3		3.1		1.45		0.80		(0.482)
Appeal to sexual practices								
Node 1	4.23		0.22		0.57		−0.00 (−0.009)	
Node 2		4.99		−0.17		0.38		0.515
Node 3		2.46		1.18		0.69		(0.509)

Note. This table presents the average item parameter values for each node, after converting each item into three pseudo-items for each questionnaire.

Correlations among latent traits from different models. To explore the relationships and differences between latent traits estimated by various models, we examined correlations among latent traits under different models in empirical data. Figure 9 presents these correlations as estimated by the six different models. For the scatter plot of the IRTree and UTree models, the distributions of latent traits appear widely dispersed, lacking a clear unified correlation direction. This suggests that the dominance and ideal point processes represent distinct cognitive decision-making mechanisms.

Notably, the agreement of reading interest (θ_1) in the ERSUTree and ORDUTree.2 models exhibit a high correlation ($r = 0.982$, $p < 0.001$), signifying consistent outcomes when employing the same unfolding model estimation for Node 1. However, in Nodes 2 and 3, when the ERSUTree model using extreme response style (η) replaces the "degree of agreement" trait (θ_2) in ORDUTree.2, it may underestimate this latent trait's actual value when θ_2 falls between −1 and 0, and overestimate it when θ_2 falls between 0 and 1, which result in biased estimations.

The relatively high correlation between ORDUTree.1 and ORDUTree.2, while not identical, hints that the presumption of either one or two target traits can lead to disparate outcomes. This underscores the existence of two slightly divergent latent traits associated with reading interest. Importantly, when we rely solely on

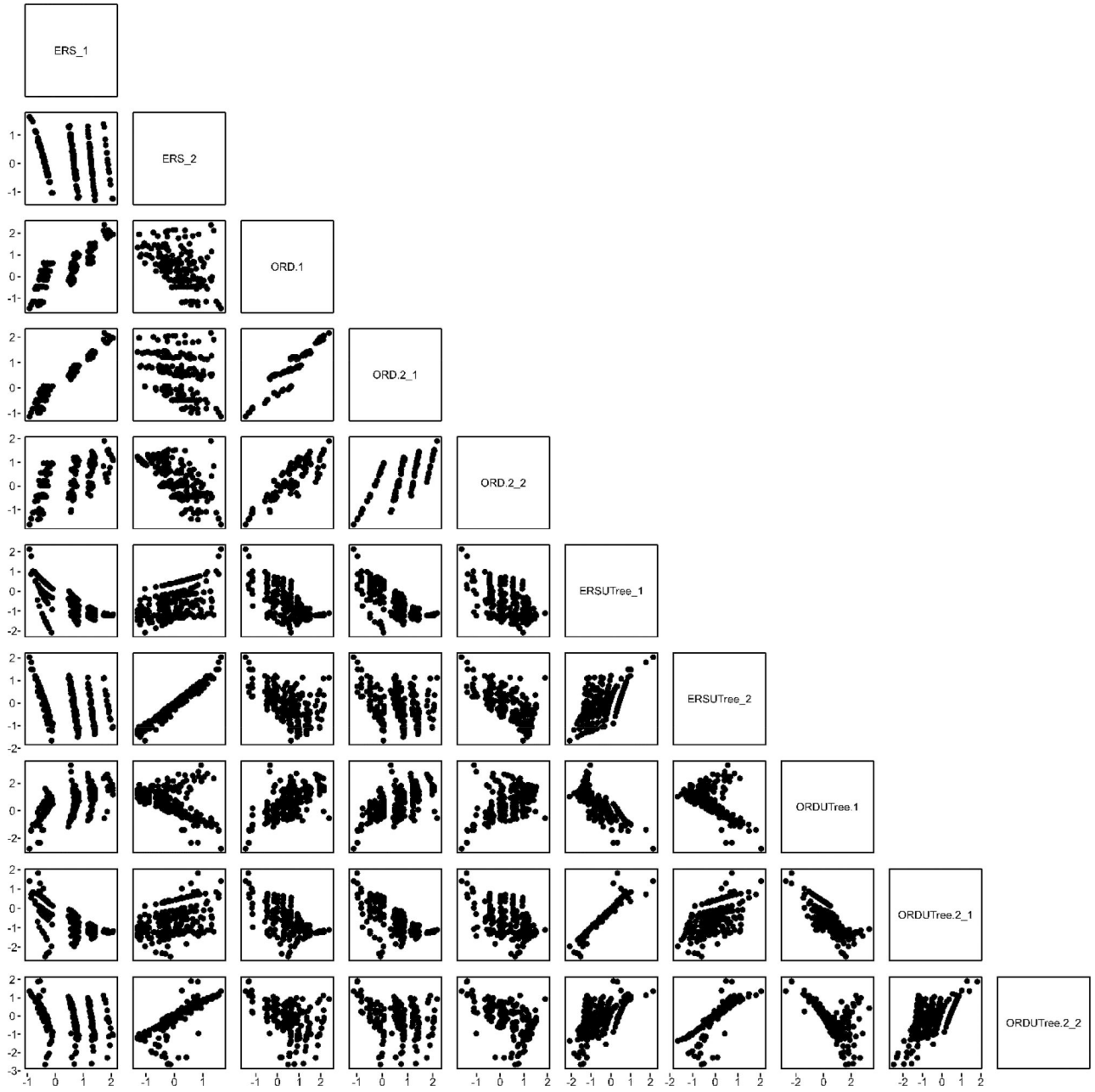

Figure 9. Scatter plots of estimated latent traits for "reading interest" questionnaires under different IRTree and UTree models. *Note.* "ERS_1" represents the latent trait from decision Stage 1 in the ERS model, which corresponds to the target trait θ, while "ERS_2" signifies the latent trait from decision Stage 2, indicating the extreme response style η. Similarly, for other models, suffixes "_1" and "_2" denote latent traits from decision Stage 1 and 2, respectively.

one latent trait to represent participants' reading interest, we may inadvertently neglect a second type of reading interest, consequently resulting in skewed estimations. The scatter plot contrasting ORDUTree.1 and ORDUTree.2 underscores that, at higher levels of the degree of agreement, outcomes derived from the ORDUTree.1 model are markedly unstable and dispersed, leading to a considerable underestimation of this particular reading interest trait.

In the ORDUTree.2 model, the moderate correlation ($r = 0.791$, $p < 0.001$) between the "agreement of reading interest" (θ_1) and the "degree of agreement" (θ_2) support the idea that these two distinct latent traits are positively related. Through scatter plots, we observe that when individuals have a higher tendency to agree with the statement at Node 1 (with θ_1 ranging between -1.5 and 0.5, thus closer to the item location parameter), the distribution of the degree of agreement (θ_2) among different individuals is quite dispersed. This dispersion indicates that individuals do not consistently show a strong tendency to particularly agree or disagree. Conversely,

when individuals tend to disagree with the statement (θ_1 being further from the item location parameter), they exhibit a clear tendency at Nodes 2 and 3 to choose options indicative of lower agreement levels, such as "strongly disagree" and "agree". Overall, these findings emphasize the unique nature of each latent trait and their relationship in different cognitive/decision processes under reading interest assessment.

Example 2: appeal of sexual practices

Datasets and analytical procedure

Subsequently, we employ a subscale from the National Health and Social Life Survey to gauge male respondents' attitudes toward sexual practices (Laumann et al., 1992). This scale consists of 15 items that assess the perceived attractiveness of various types of sexual practices. Participants were asked to select the most fitting option from four Likert items: 0 = not at all appealing, 1 = not appealing, 2 = somewhat appealing, and 3 = very appealing. Given the neutral nature of the descriptions in these items, there is no need for reverse scoring. Moreover, this neutrality likely leads respondents with moderate sexual attitudes to agree with these items, suggesting that considering an ideal point response process might be appropriate for this data set (Jin et al., 2022). After data cleansing and addressing missing values, the dataset used for formal analysis encompassed 1,397 respondents' data. The simulation study has confirmed that both IRTree and UTree models could effectively estimate item parameters and discern the decision processes with scales exceeding 10 items, even with small sample sizes. Thus, a 15-item scale with responses from 1397 participants is deemed suitable for analysis. The procedures and metrics employed are consistent with those used in Example 1.

Results

Fit indices. The findings mirror those from the reading interest analysis. As shown in Table 5 and Figure 7, the UTree model has lower AIC and SABIC and higher h^2 values than for the IRTree model. This reaffirms that respondents, when answering this Likert questionnaire, adhere to the ideal point process rather than the dominance process, suggesting that the UTree model is more apt for analyzing Likert items. The ORDUTree.2 model, with the smallest AIC, SABIC and highest h^2 values, again emerges as the superior model. This solidifies the idea that respondents' answers to this Likert scale were driven by two distinct types of target traits—agreement tendency (Node 1) and degree of agreement (Node 2 and 3) toward sexual practices—rather than by a single target trait or a blend of target trait and extreme response style. This finding also indicates that when respondents answer Likert scales, they choose extreme statement options based on target traits rather than extreme response style. Moreover, these traits differ between the two decision stages.

Item parameters and latent traits. We conduct further analysis on the item parameters and latent trait estimations obtained from the best-fitting ORDUTree.2 model, as shown in Table 6. Most items exhibit high levels of discrimination across all three nodes. In Node 1, respondents' agreement with the appeal of sexual practices trait landed at a moderate level ($\theta_1 = -0.00$). The proximity of this trait to the item and threshold parameters implies that respondents generally agree with the statements about the appeal of sexual practices, suggesting a higher degree of approval for such practices.

Within Node 2, the degree of respondents' agreement ($\theta_2 = 0.559$) closely align with the item and threshold parameters. This alignment suggests that respondents were more inclined to opt for "not appealing" rather than "not at all appealing," which points to a stronger agreement regarding sexual practices. In contrast, Node 3 reveals a more pronounced difference between the respondents' traits and the item and threshold parameters. This divergence suggests that participants were more likely to rate practices as "somewhat appealing" than "very appealing". This nuance points to a relatively strong, though not maximal, agreement strength. Overall, the items within the appeal of sexual practices scale demonstrate strong discrimination, and respondents' inclination toward the appeal of sexual practices was moderately elevated.

Correlations among latent traits from different models. Figure 10 showcases the correlations of latent traits derived from the IRTree and UTree models. Mirroring the observations from the reading interest analysis, the scatter plots representing latent traits across both IRTree and UTree models do not display a distinct directional trend. More notably, a number of scatter plots that delineate latent traits as estimated by IRTree and UTree models display nonlinear associations. This observation reinforces the idea that dominance and ideal point processes emerge from separate cognitive decision-making mechanisms.

Given the shared model type at Node 1 between the ERSUTree and ORDUTree.2 models, it is unsurprising to observe a relatively higher correlation in

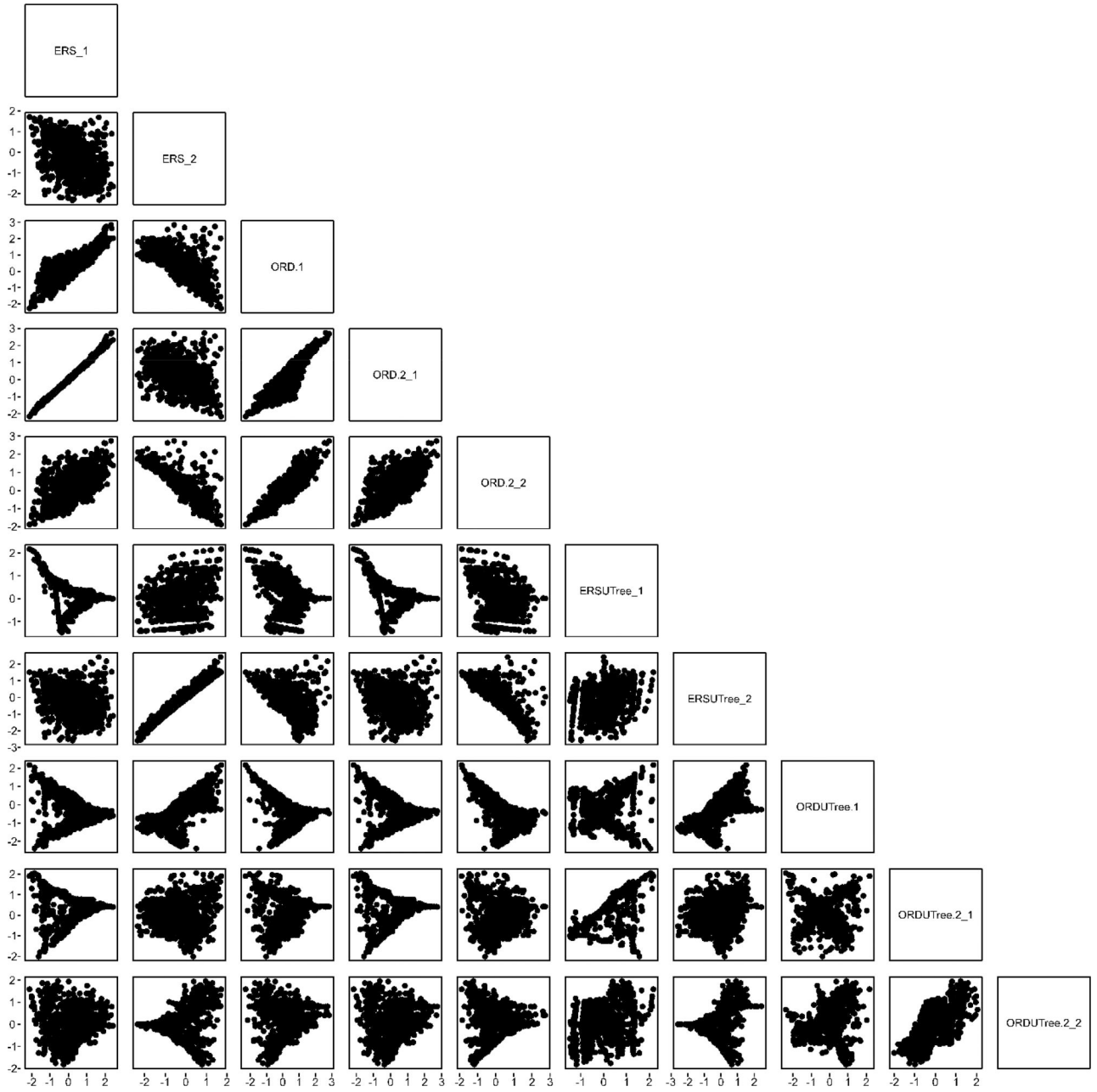

Figure 10. Scatter plots of estimated latent traits for "appealing to sexual practices" questionnaires under different IRTree and UTree models.

their θ_1 estimates. However, as respondents' "agreement with the appeal of sexual practices" intensifies, the ERSUTree model's estimated results become increasingly volatile. The bias becomes more pronounced, especially when θ_1 approaches 1.5. In Nodes 2 and 3, there is a noticeable discrepancy between the ERSUTree's estimates of ERS (η) and the actual "degree of agreement" (θ_2) of the respondents. This discrepancy is particularly salient when θ_2 reaches higher levels, leading to scattered estimation outcomes and pronounced bias. Incorrectly using this model significantly compromises both the accuracy and interpretability of the derived results.

In the ORDUTree.2 model, the "agreement of the appeal of sexual practices" (θ_1) and "degree of agreement" (θ_2) show a moderate correlation ($r = 0.684, p < 0.001 p < 0.001$), indicating a positive association and a moderate level of correlation between these two latent traits. Scatter plots reveal that regardless of whether individuals' target trait levels at Node 1 are close to the item position parameters, the distribution of their target trait levels at

Nodes 2 and 3 is quite scattered. This indicates that regardless of whether individuals agree with the statement at Node 1, there is no consistency in their choices at specific options in subsequent nodes. The two different target traits across different decision-making stages exhibit significant variability, reflecting their distinctiveness in influencing decision outcomes.

Meanwhile, the low correlation between ORDUTree.1 and ORDUTree.2 suggests that using a single target trait for estimation could lead to significantly disparate results. In Node 1, when respondents' "agreement of the appeal of sexual practices" trait is either low or high, the results estimated using ORDUTree.1 are highly dispersed and unstable, leading to substantial estimation bias. In Nodes 2 and 3, when respondents' "degree of agreement" is at higher levels, the ORDUTree.1 model estimates also suffer from extreme dispersion and instability, indicating that using only one latent target trait is not feasible and would result in significant estimation bias.

Conclusions and discussion

Conclusions

Building upon the existing IRTree models, in this study, we considered the ideal point process-based unfolding model to restructure three distinct types of Unfolding Tree models. This approach offers a novel perspective and a versatile modeling framework, offering a new viewpoint on the relationship between latent traits and the distance to the ideal point to reflect respondents' multi-stage ideal point responses. This enables a deeper exploration into Likert scale responses and the specific latent traits driving decision-making.

Our simulation study validated that the fit indices can accurately discern the true model that aligns with the data's decision-making process. Moreover, when the estimated model aligns correctly, both the IRTree and UTree models showcase satisfactory performance. However, errors arise when there is a mismatch. Whether it is mistakenly applying the IRTree model to UTree data, erroneously attributing target traits to ERS, or forcibly estimating multiple traits as a single trait, substantial biases are introduced to individual parameter estimates. These findings not only affirm the fit indices' and models' viability but also underscore the critical necessity of employing the correct model in data analysis.

Lastly, by examining two concrete instances, we contrasted all the IRTree and UTree models. Results revealed that respondents are more likely to undertake a two-stage, three-node decision based on the ideal point process. Crucially, distinct decision stages are underpinned by different target traits. This deepens our comprehension of the intricate decision-making mechanisms that respondents deploy when engaging with Likert scales. Consequently, it offers invaluable empirical evidence that enriches our grasp of the underlying cognitive processes and latent traits influencing their decisions.

Discussion

Here we discuss the specific results and findings from our research. In the simulation study, first and foremost, we discovered that both AIC and SABIC effectively identify the true (correct) model underpinning response data across various conditions. This validates the applicability of these two fit indices when analyzing IRT and unfolding-related models, which is consistent with existing studies (De Boeck & Partchev, 2012). Interestingly, when respondents answer based on a single latent trait, irrespective of whether it is rooted in the ideal point or dominance process, both ERSUTree and ORDUTree.2 exhibit high sensitivity. They can directly report that the sigma matrix contains negative eigenvalues during the estimation process. This immediate feedback eliminates the need for further inspection of fit indices and item parameters. Consequently, it offers a direct indication that respondents are not basing their answers on two distinct latent traits. This greatly prevents the waste of resources in unnecessary comparisons and the biases that arise from using incorrect models.

We found that when the estimated model is correctly identified, all IRTree and UTree models consistently return accurate item and individual parameters. Interestingly, as the number of items and sample size increase, the precision of these estimations also improves, a finding in line with earlier research (Roberts & Laughlin, 1996). The correlation between latent traits has minimal impact on model estimation performance, indicating that the model can aptly estimate the relationships between various latent traits.

However, when the estimated model is inaccurate, it can lead to significant biases in the estimation of respondents' latent traits. For instance, when respondents select extreme response options based on target traits, mistakenly classifying these target traits as ERS can induce substantial estimation biases in person parameters. This bias becomes especially pronounced when respondents decide based on the ideal point process, and the ERSUTree model is incorrectly

employed. Empirical results echo this observation. When respondents might be operating on an ideal point process, and make decisions in different stages based on two distinct target traits, using the ERSUTree model produces unstable and dispersed biased estimates for those with either high or low true trait levels. This could fundamentally misconstrue the concept of the respondent's true trait level.

Furthermore, when respondents decide based on different target traits in various stages, inaccurately assuming it is rooted in a singular target trait can also cause a systemic shift in person parameters, leading to significant biases across most parameters. Empirical studies also affirm this, when respondents possibly decide based on an ideal point process and two different types of target traits, using the ORDUTree.1 model results in consistently biased estimates for target traits, particularly when the second target trait θ_2 is high.

Thus, effectively and judiciously employing fit indices to discern the potential genuine decision-making process and latent traits behind response data can substantially mitigate the biases that arise from using inappropriate models. This approach fosters a more accurate derivation of item parameters and individual latent trait levels.

In the empirical study, we observed that all UTree models outperformed the IRTree models. This further substantiates the notion that when respondents answered Likert scales, they might not be adhering to the dominance response process but rather the unfolding response process (Thurstone, 1928). This observation aligns with numerous existing research findings (Chernyshenko et al., 2001; 2007). In essence, when respondents engage with Likert scales, they are most likely to respond affirmatively only when the item's phrasing closely aligns with their latent trait level.

Next, among the UTree models, the ORDUTree.2 model demonstrated the best fit, indicating that respondents were making decisions based on target traits rather than extreme response style when choosing whether to agree with the extreme expression option. While employing the ERSUTree model yielded a better fit than the traditional IRTree model at this juncture, it is perilous to simply attribute respondents' specific option choices to extreme response style (Jin et al., 2022; Li et al., 2025). Respondents might be choosing specific options based on target traits rather than extreme response styles. The superior fit of the ERSUTree model might only result from the application of the unfolding model during Stage 1 decision-making. Hence, it is imperative for us to further compare the ERSUTree and ORDUTree models to identify a more accurate model that genuinely reflects the respondents' latent traits. Upon further analysis, the superior performance of ORDUTree.2 signifies that respondents are not choosing extreme options based on extreme response style. Instead, such choices arise because the item phrasing significantly diverges from their inherent trait level. Using the ERSUTree model in such a scenario would mistakenly interpret target traits as extreme response styles, introducing a conceptual bias in our understanding of the respondents' latent traits. This would gravely mislead our interpretation of the underlying latent traits driving the response process.

Finally, the superior performance of the ORDUTree.2 model over ORDUTree.1 underscores that respondents, during two distinct decision stages, base their choices on different target traits instead of a single one. Scatter plots from two real data applications of the ORDUTree.2 model demonstrate that in many instances, regardless of whether individuals agree or disagree with the statement at Stage 1, there is no consistency in their choices at specific options during Stage 2. This indicates significant variability between the two different target traits across different decision-making stages, providing indirect evidence for the existence of distinct target traits for agreement and strength of agreement. This observation is consistent with prior research on IRTree models (Jeon et al., 2017). Moreover, in comparison to ORDUTree.1, ORDUTree.2 offers enhanced flexibility. By sidestepping potential estimation biases that arise from contradictory assumptions about target trait levels at nodes 1 and nodes 2,3, the ORDUTree.2 model can more precisely estimate target trait parameters in empirical data.

In summary, this research offers a flexible estimation framework based on the ideal point process for analyzing responses to Likert scales. Based on our findings, we recommend prioritizing the use of the ORDUTree.2 model when estimating on Likert scales. This approach allows for the free estimation of potential distinct target traits during different decision stages, resulting in a more accurate retrieval of item parameters and respondents' latent trait levels. Alternatively, as a comprehensive approach, one can employ both AIC and SABIC to evaluate each of the three UTree models, then proceed with the one that demonstrates the optimal fit. This rigorous approach to model selection ensures thoroughness, though researchers should anticipate the potential for

increased time and resource commitments, especially with large datasets.

Limitations and future directions

This study also has some limitations. Firstly, our research mainly employs a four-point scoring scale for demonstration. Future research should utilize Likert scales encompassing a wider range of scores, supplemented with more empirical data, to validate the model's efficacy further and to provide deeper insights into respondents' underlying decision-making processes. Secondly, the primary emphasis of this study is on the extreme response style. It would be advantageous for subsequent research to extend the application of UTree models to explore different response styles, such as the Midpoint Response Style and Acquiescence Response Style. Thirdly, this study aims to establish multi-process response models based on the ideal point process, using only unidimensional latent traits (ERS or target traits) at each node. However, recent studies in the IRTree field are increasingly developing multidimensional or mixture model hypotheses that simultaneously consider ERS and target traits (Alagöz & Meiser, 2023; Kim & Bolt, 2021; Merhof & Meiser, 2023). These models may better reflect respondents' actual conditions and have shown good psychometric performance. Therefore, future UTree model development could further explore multidimensional traits or mixture models, which are necessary and meaningful. Finally, this study primarily uses Likert datasets to analyze the performance and implications of the UTree model, without incorporating other external criterion variables. Future research could further explore the latent traits derived from these UTree models, particularly the ORDUTree.2 model, in relation to external criterion variables. This would provide more evidence for further examining and validating the substantive values of these traits.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant 202206040119 from the China Scholarship Council.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alagöz, O. E. C., & Meiser, T. (2024). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement*, *84*(5), 957–993. https://doi.org/10.1177/00131644231206765

Andrich, D. (1988). The application of an unfolding model of the pirt type to the measurement of attitude. *Applied Psychological Measurement*, *12*(1), 33–51. https://doi.org/10.1177/014662168801200105

Andrich, D., & Luo, G. Z. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*(3), 253–276. https://doi.org/10.1177/014662169301700307

Arce-Ferrer, A. J., & Ketterer, J. J. (2003). The effect of scale tailoring for cross-cultural application on scale reliability and construct validity. *Educational and Psychological Measurement*, *63*(3), 484–501. https://doi.org/10.1177/0013164403063003009

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*(6), 1235–1245. https://doi.org/10.1016/j.paid.2005.10.018

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*(2), 171–187. https://doi.org/10.1093/pan/mpi010

Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 608–628. https://doi.org/10.1207/S15328007SEM0704_5

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69–83. https://doi.org/10.1037/met0000106

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi—process IRT models: A review and tutorial. *The British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods*, *18*(2), 252–275. https://doi.org/10.1177/10944281145559

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, Q., Luo, W., Palardy, G. J., Glaman, R., & McEnturff, A. (2017). The efficacy of common fit indices for enumerating classes in growth mixture models when nested data structure is ignored: A Monte Carlo study. *Sage Open*, *7*(1), 1–19. https://doi.org/10.1177/2158244017700459

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562. https://doi.org/10.1207/S15327906MBR3604_03

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*(1), 88–106. https://doi.org/10.1037/1040-3590.19.1.88

Clarke, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review*, *18*(3), 301–324. https://doi.org/10.1108/02651330110396488

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, *57*(3), 145–158. https://doi.org/10.1037/h0060984

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(Code Snippet 1), 1–28. https://doi.org/10.18637/jss.v048.c01

de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, *30*(3), 216–232. https://doi.org/10.1177/0146621605282772

Drasgow, F., Chernyshenko, O., & Stark, S. (2010). 75 years after Likert: Thurstone was right!. *Industrial and Organizational Psychology*, *3*(4), 465–476. https://doi.org/10.1111/j.1754-9434.2010.01273.x

Duck-Mayr, J., & Montgomery, J. (2023). Ends against the middle: Measuring latent traits when opposites respond the same way for antithetical reasons. *Political Analysis*, *31*(4), 606–625. https://doi.org/10.1017/pan.2022.33

Enders, C., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 75–95. https://doi.org/10.1080/10705510701758281

Fang, Y. (2020). *The applicability research of GGUM in different response categories of personality test* [Master, East China Normal University]. CNKI: Master http://kns.cnki.net/KCMS/detail/detail.aspx?FileName=1020637076.nh&DbName=CMFDTEMP

Greenleaf, E. A. (1992). Improving rating-scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*(2), 176–188. https://doi.org/10.2307/3172568

Guo, Q., Miao, J., & Wang, W. (2006). Unfolding IRT Model and the Non-cumulative Response Mechanism in Personality Tests. *Psychological Exploration*, (01), 66–69.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement*, *15*(2), 153–169. https://doi.org/10.1177/014662169101500205

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309. https://doi.org/10.1177/0022022189203004

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

Jeon, M., & De Boeck, P. (2019a). An analysis of an item response strategy based on knowledge retrieval. *Behavior Research Methods*, *51*(2), 697–719. https://doi.org/10.3758/s13428-018-1064-1

Jeon, M., & De Boeck, P. (2019b). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *The British Journal of Mathematical and Statistical Psychology*, *72*(3), 517–537. https://doi.org/10.1111/bmsp.12182

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, *42*(4), 467–490. https://doi.org/10.3102/1076998616688015

Jin, K., Wu, Y., & Chen, H. (2022). A new multiprocess IRT model with ideal points for Likert-type items. *Journal of Educational and Behavioral Statistics*, *47*(3), 297–321. https://doi.org/10.3102/10769986211057160

Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*(1), 92–114. https://doi.org/10.3102/1076998609340529

Kim, N., & Bolt, D. M. (2021). A mixture irtree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*(1), 131–154. https://doi.org/10.1177/0013164420913915

LaHuis, D. M., Blackmore, C. E., Bryant-Lees, K. B., & Delgado, K. (2019). Applying item response trees to personality data in the selection context. *Organizational*

Research Methods, *22*(4), 1007–1018. https://doi.org/10.1177/1094428118780310

Laumann, E. O., Gagnon, J. H., Michael, R. T., & Michaels, S. (1992). *National health and social life survey*. University of Chicago and National Opinion Research Center. [Producer]. Inter-University Consortium for Political and Social Research [Distributor]. https://doi.org/10.3886/ICPSR06647.v2

Li, Z., Li, L., Zhang, B., Cao, M., & Tay, L. (2025). Killing two birds with one stone: accounting for unfolding item response process and response styles using unfolding item response tree models. *Multivariate Behavioral Research*, *60*(2), 161–183. https://doi.org/10.1080/00273171.2024.2394607

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1–55.

Lord, F. (1952). A theory of test scores. *Psychometric Monographs*, (7), 1–84.

Luo, G. Z. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology*, *45*(2), 224–248. https://doi.org/10.1006/jmps.2000.1310

Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, *88*(4), 1354–1380. https://doi.org/10.1007/s11336-023-09901-0

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality and Quantity*, *37*(3), 277–302. https://doi.org/10.1023/A:1024472110002

National Center for Education Statistics (2008). PISA attitude analysis study. Internal Report.

OECD (2018). *Programme for International Student Assessment 2018 Database* https://www.oecd.org/pisa/data/2018database/

Park, M., & Wu, A. D. (2019). Item response tree models to investigate acquiescence and extreme response styles in likert-type rating scales. *Educational and Psychological Measurement*, *79*(5), 911–930. https://doi.org/10.1177/0013164419829855

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes*. Academic Press. https://doi.org/10.1016/b978-0-12-590241-0.50006-x

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3–32. https://doi.org/10.1177/01466216000241001

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, *20*(3), 231–255. https://doi.org/10.1177/014662169602000305

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464., MR 0468014. https://doi.org/10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343. https://doi.org/10.1007/BF02294360

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *The Journal of Applied Psychology*, *91*(1), 25–39. https://doi.org/10.1037/0021-9010.91.1.25

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions* [PhD University of Oxford].

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response 136 process. *Organizational Research Methods*, *15*(3), 363–384. https://doi.org/10.1177/109442811243970

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *The Journal of Applied Psychology*, *94*(5), 1287–1304. https://doi.org/10.1037/a0015899

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–554. https://doi.org/10.1086/214483

Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, *50*(6), 2325–2344. https://doi.org/10.3758/s13428-017-0997-0

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*(2), 178–189. https://doi.org/10.1016/j.jrp.2012.10.010

Zeng, B., Jeon, M., & Wen, H. (2024). How does item wording affect participants' responses in Likert scale? Evidence from IRT analysis. *Frontiers in Psychology*, *15*, 1304870. https://doi.org/10.3389/fpsyg.2024.1304870

Zeng, B., Wen, H., & Zhang, J. (2020). How does the valence of wording affect features of a scale? The method effects in the Undergraduate Learning Burnout scale. *Frontiers in Psychology*, (*11*), 585179. https://doi.org/10.3389/fpsyg.2020.585179

Appendix A

Table A1. Average BIC value and the percentage of times each model exhibits the lowest fit index values across all conditions for six models.

		Data Generation Model					
Fitted model	Index	ERS	ORD.1	ORD.2	ERS UTree	ORD. UTree.1	ORD. UTree.2
ERS	BIC	**31045.59 (100%)**	20943.92 (0%)	32066.78 (0%)	31862.47 (11%)	34668.53 (1%)	34955.14 (0%)
ORD.1	BIC	32547.35 (0%)	**20303.75 (100%)**	31979.25 (0%)	33100.89 (0%)	34454.03 (5%)	35030.17 (0%)
ORD.2	BIC	32092.17 (0%)	20511.50 (0%)	**31023.68 (100%)**	32785.58 (0%)	34444.91 (1%)	34764.49 (9%)
ERSUTree	BIC	31147.44 (0%)	–	32171.50 (0%)	**31397.45 (89%)**	–	34514.38 (1%)
ORDUTree.1	BIC	31704.80 (0%)	–	31516.78 (0%)	32152.53 (0%)	**33203.96 (93%)**	34509.20 (0%)
ORDUTree.2	BIC	31323.30 (0%)	–	31334.43 (0%)	31583.42 (0%)	–	**33896.44 (89%)**

Note. Each condition was replicated 100 times (consistent with the subsequent simulation analysis).

Appendix B

Table B1. Average Bias and RMSE for data generated and estimated with ERS model across different conditions.

		Cor = 0		Cor = 0.3		Cor = 0.6	
N	*I*	Bias	RMSE	Bias	RMSE	Bias	RMSE
500	5	0.00	0.19	−0.01	0.18	−0.01	0.19
500	10	0.00	0.16	−0.01	0.17	−0.01	0.16
500	20	−0.01	0.16	−0.02	0.16	−0.02	0.16
1000	5	0.00	0.14	0.00	0.14	−0.01	0.14
1000	10	0.00	0.12	−0.01	0.12	−0.01	0.13
1000	20	−0.01	0.11	−0.01	0.11	−0.01	0.11
2000	5	0.01	0.10	0.00	0.10	0.00	0.10
2000	10	0.00	0.09	0.00	0.09	0.00	0.09
2000	20	0.00	0.08	−0.01	0.08	−0.01	0.08

Table B2. Average Bias and RMSE for data generated and estimated with ORD.1 and ORDUTree.1 model across different conditions.

		ORD.1		ORDUTree.1	
N	*I*	Bias	RMSE	Bias	RMSE
500	5	−0.02	0.23	−0.05	0.34
500	10	−0.02	0.20	−0.01	0.22
500	20	−0.03	0.21	0.00	0.20
1000	5	−0.01	0.17	−0.01	0.24
1000	10	−0.01	0.16	0.00	0.18
1000	20	–	–	0.00	0.16
2000	5	−0.01	0.12	0.00	0.18
2000	10	0.00	0.11	0.00	0.13
2000	20	–	–	0.00	0.11

Table B3. Average Bias and RMSE for data generated and estimated with ORD.2 model across different conditions.

		Cor = 0		Cor = 0.3		Cor = 0.6	
N	*I*	Bias	RMSE	Bias	RMSE	Bias	RMSE
500	5	0.00	0.18	−0.01	0.19	−0.02	0.19
500	10	−0.01	0.16	−0.01	0.17	−0.01	0.16
500	20	−0.01	0.15	−0.02	0.16	−0.02	0.16
1000	5	0.01	0.15	−0.01	0.14	−0.01	0.14
1000	10	0.00	0.12	−0.01	0.12	−0.01	0.13
1000	20	0.00	0.11	−0.01	0.11	−0.01	0.12
2000	5	0.00	0.10	0.00	0.10	0.00	0.11
2000	10	0.00	0.09	0.00	0.09	0.00	0.09
2000	20	0.00	0.08	0.00	0.08	−0.01	0.08

Table B4. Average Bias and RMSE for data generated and estimated with ERSUTree model across different conditions.

		Cor = 0		Cor = 0.3		Cor = 0.6	
N	*I*	Bias	RMSE	Bias	RMSE	Bias	RMSE
500	5	0.02	0.26	0.00	0.26	−0.01	0.23
500	10	0.00	0.21	−0.01	0.20	−0.01	0.19
500	20	0.00	0.17	−0.01	0.17	−0.01	0.17
1000	5	0.01	0.21	0.00	0.20	0.00	0.18
1000	10	0.00	0.15	0.00	0.15	−0.01	0.14
1000	20	0.00	0.13	−0.01	0.13	0.00	0.13
2000	5	0.00	0.16	0.00	0.15	0.00	0.13
2000	10	0.00	0.11	0.00	0.11	0.00	0.10
2000	20	0.00	0.09	0.00	0.09	0.00	0.09

Table B5. Average Bias and RMSE for data generated and estimated with ORDUTree.2 model across different conditions.

		Cor = 0		Cor = 0.3		Cor = 0.6	
N	*I*	Bias	RMSE	Bias	RMSE	Bias	RMSE
500	5	0.02	0.29	0.01	0.28	0.00	0.26
500	10	0.01	0.22	0.00	0.21	0.00	0.21
500	20	0.00	0.19	0.00	0.18	0.00	0.19
1000	5	0.01	0.25	0.01	0.23	0.00	0.21
1000	10	0.01	0.18	0.00	0.17	0.00	0.16
1000	20	0.00	0.15	0.00	0.14	0.00	0.15
2000	5	0.01	0.21	0.00	0.18	−0.01	0.17
2000	10	0.00	0.13	0.00	0.13	0.00	0.12
2000	20	0.00	0.11	0.00	0.11	0.00	0.10

Appendix C

Table C1. Average Bias and RMSE of each item parameter for data generated from six models across all conditions.

	ERS	ORD.1	ORD.2	ERS UTree	ORD.UTree.1	ORD.UTree.2
Bias						
α_1	−0.01	−0.02	−0.02	0.03	0.02	0.03
α_2	−0.03	−0.04	−0.03	−0.03	0.03	0.05
$\delta_1(\beta_1)$	0.00	0.00	0.00	0.00	0.00	0.00
$\delta_2(\beta_2)$	0.00	0.00	0.00	0.00	0.00	0.00
τ_1	–	–	–	−0.03	−0.04	−0.03
τ_2	–	–	–	–	−0.06	−0.05
Cor	0.01	–	0.01	0.02	–	0.03
RMSE						
α_1	0.14	0.14	0.14	0.20	0.17	0.20
α_2	0.21	0.22	0.21	0.21	0.23	0.26
$\delta_1(\beta_1)$	0.10	0.11	0.10	0.18	0.18	0.19
$\delta_2(\beta_2)$	0.17	0.22	0.17	0.19	0.25	0.25
τ_1	–	–	–	0.14	0.14	0.15
τ_2	–	–	–	–	0.20	0.19
Cor	0.03	–	0.03	0.04	–	0.05

Note. α_1, $\delta_1(\beta_1)$, τ_1 represent the discrimination, item location (difficulty), and threshold parameters at the first decision stage, i.e., Node 1. α_2, $\delta_2(\beta_2)$, τ_2 represent the discrimination, item location (difficulty), and threshold parameters for the second decision stage, i.e., Nodes 2 and 3. Given that these two nodes belong to the same decision stage and utilize the same item location (difficulty) parameters, we have, for the sake of concise representation, combined the distinct parameters of Nodes 2 and 3 from the original text, forming an integrated item parameter representing the second decision stage.

Appendix D

Table D1. Average Absolute Bias and RMSE of estimated θ and η in six models across all conditions and replications.

	Data Generation Model									
	ERS		ORD.1	ORD.2		ERS UTree		ORD UTree.1	ORD UTree.2	
Fitted model	θ	η	θ	θ_1	θ_2	θ	η	θ	θ_1	θ_2
Absolute bias										
ERS	**0.42**	**0.42**	0.45	0.42	0.57	0.54	0.42	0.8	0.55	0.45
ORD.1	0.41	0.43	**0.36**	–	–	–	–	0.53	–	–
ORD.2	0.42	0.57	0.59	**0.42**	**0.42**	0.59	0.58	0.8	0.71	0.68
ERSUTree	0.69	0.44	0.75	0.7	0.58	**0.46**	**0.44**	0.72	0.46	0.45
ORDUTree.1	0.64	0.63	0.72	0.71	0.73	–	–	**0.37**	–	–
ORDUTree.2	0.70	0.68	0.79	0.71	0.75	0.47	0.66	0.65	**0.46**	**0.47**
RMSE										
ERS	**0.53**	**0.53**	0.57	0.53	0.74	0.69	0.53	1.04	0.7	0.57
ORD.1	0.52	0.54	**0.46**	–	–	–	–	0.7	–	–
ORD.2	0.53	0.74	0.74	**0.53**	**0.53**	0.76	0.75	1.04	0.9	0.87
ERSUTree	0.92	0.56	0.98	0.94	0.74	**0.62**	**0.55**	0.96	0.62	0.57
ORDUTree.1	0.87	0.85	0.99	0.97	0.99	–	–	**0.5**	–	–
ORDUTree.2	0.94	0.89	1.03	0.95	1.01	0.64	0.87	0.86	**0.63**	**0.63**

Note. Since the Bias of the estimated θ and η from these models is less than 0.0001, reporting this result is not particularly meaningful. Therefore, we have opted to report the Absolute Bias in this section.