Psychometric Modeling Approaches for Understanding Item Response Process in Cognitive and Noncognitive Assessments

by

Nana Kim

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Educational Psychology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: 06/08/22

The dissertation is approved by the following members of the Final Oral Committee:

Daniel M. Bolt, Professor, Educational Psychology
James Wollack, Professor, Educational Psychology
Jee-Seon Kim, Professor, Educational Psychology
James Pustejovsky, Associate Professor, Educational Psychology
Karl Rohe, Associate Professor, Statistics

Table of Contents

List of Tables

List of Figures

Abstract

In educational and psychological assessments, attending to item response process can be useful in understanding and improving the validity of measurement. This dissertation consists of three studies each of which proposes and applies item response theory (IRT) methods for modeling and understanding cognitive/psychological response process in assessment. The first study presents a noncompensatory multidimensional IRT model that reflects underlying components for solving passage-based items in reading comprehension tests. The second study proposes a mixture item response tree (IRTree) model that accommodates the possibility of a mixture of respondents exhibiting different response processes in responding to rating scale items. The third study examines systematic differences in response process across fast and slow item responses by attending to both content trait and response style and their possible differential influences across response types. Each study introduces different modeling approaches for incorporating item response process and discusses their practical implications. The dissertation as a whole contributes to understanding item response process in both cognitive and noncognitive assessments.

Acknowledgments

I would like to take a moment to thank the people who have helped and supported me during my Ph.D. journey. This journey would not have been possible without them. First, I would like to express my sincere gratitude to my advisor, Daniel Bolt. I have been fortunate to have him as my advisor and mentor. His expertise and enthusiasm in this field has inspired me and his caring mentorship has kept me along the way. I am truly thankful to him for generously sharing his time and insights with me. His continuous support, encouragement, guidance, inspiration, and patience have been instrumental to my development as a researcher.

I am also grateful to my dissertation committee members, James Wollack, Jee-Seon Kim, James Pustejovsky, and Karl Rohe, for their encouragement and many insightful comments and suggestions. I also want to thank my wonderful professors in Quantitative Methods area, including David Kaplan, for their encouragement and valuable advice that have been a great support throughout the past few years. I would like to extend my thanks to my colleagues in Educational Psychology department and friends I met in Madison. They have enriched my Ph.D. journey and Madison life in many ways. It has always been a joy to spend time and share memorable moments with them. I must also thank my friends who provided support and friendship no matter the distance.

Finally, I would like to express my deepest gratitude to my family. I am deeply indebted to my parents, Jinmo Kim and Kyungsuk Lee, who have taught me the art of living, perseverance, wisdom, and kindness. I am sincerely grateful for their unwavering support and unconditional love through my entire life. Many thanks also go to my brother, Baro Kim, for his support and encouragement. I would not have made it this far without their love, belief, and support. I am also very much thankful to my amazing husband, Kyusic Park, for his constant support and love. I am lucky to have met him in Madison and to have him in my life. I truly thank him for always being by my side through good and bad times.

1 Introduction

Measurement is without question a fundamental and crucial component in education and psychology. Through applications of measurement models, we are able to understand individual differences in attributes, abilities, and experiences through latent constructs, such as personality or intellectual proficiencies. Assessments have particularly provided stakeholders in education with meaningful bases for evaluation and decision-making in improving education policy and practice. For instance, schools and teachers regularly use assessment results to diagnose student learning strengths and weaknesses, monitor student learning progress, and plan interventions and remediations. Also, educators and policymakers have used large-scale assessment data to evaluate education policy, assess the quality and effectiveness of education systems, and identify factors promoting high-quality education.

Assessments can be distinguished into a cognitive or noncognitive assessment depending on which constructs are being measured. Cognitive assessments measure intellectual skills or cognitive ability constructs, such as literacy, numeracy, information processing, reasoning, and problem-solving skills. On the other hand, noncognitive assessments in education measure behaviors, attitudes, and mindsets related to socio-emotional learning constructs and learning experiences. While cognitive assessments have had a long history in education systems, noncognitive assessments have only recently become of growing interest and importance due to the increased emphasis on noncognitive aspects of learning in education. As both cognitive and noncognitive constructs are acknowledged as important factors contributing to educational and life success, many national and international assessments such as the

Programme for International Student Achievement (PISA), Trends in International Mathematics and Science Study (TIMSS), and National Assessment of Educational Progress (NAEP) nowadays measure both construct types. Measuring both forms of constructs can help us achieve a comprehensive understanding of student learning and development.

In both cognitive and noncognitive assessments, it is often useful to attend to the psychological processes that may be involved in item responses. Though it is possible to measure latent constructs without understanding the underlying item response process, we can improve the measurement of constructs and better inform about respondents and items by attending to how individual respondents arrived at final item responses. Measurement models incorporating item response process can provide a basis for more informative feedback (e.g., which aspect(s) of items a respondent is having difficulty solving) and for improving the validity of assessments by helping us understand what is really being measured and whether any unintended process is interfering with the measurement of intended-to-be-measured construct. In this respect, there has been a growing interest in understanding and modeling item response processes, especially with the recent advancement in statistical computations and increased use of computer-based tests. More recent attention has also focused on the importance of understanding response process for interpretation of latent metrics used for scoring and the measurement of score gains (Bolt & Liao, 2022).

This dissertation comprises three studies examining psychometric modeling approaches that allow a better reflection of the underlying cognitive/psychological response process and/ or behaviors that individuals may exhibit when responding to items in cognitive or noncognitive assessments. The first study focuses on cognitive assessments, specifically reading comprehension tests, while the latter two studies focus on noncognitive assessments in which rating scales are used. Each study provides a critical view on existing approaches of modeling item responses, and proposes and evaluates alternative modeling approaches that can improve our understanding of item response process.

The three studies are each presented as a chapter. In Chapter 2, I propose a noncompensatory item response theory (IRT) model that is sensitive to the response processes involved in reading comprehension tests where test items are organized around passages. In contrast to compensatory models commonly applied to passage-based test structure (e.g., testlet and bifactor models), this model incorporates a noncompensatory interaction between two underlying components to solving passage-based items, and thus, provides a potentially psychologically meaningful representation of item response processes that goes beyond simply accounting for statistical dependence among items within passages. My study evaluates the model in comparison to a bifactor model, and further demonstrates the practical implication and usefulness of the model through a real data application.

In Chapter 3, I investigate response behavior heterogeneity across individuals by proposing a mixture item response tree (IRTree) model that incorporates the possibility of a mixture of respondents following different underlying response processes in the use of a rating scale to respond to noncognitive test items. The model specifically accounts for a mixture of respondents exhibiting and not exhibiting response styles, namely, content-irrelevant tendencies to select particular response categories (Paulhus, 1991). The study provides critical insight into IRTree models (Böckenholt, 2012; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016) which have been promoted as an intuitive and flexible statistical framework to model item response processes. Through simulation and empirical studies, I demonstrate the presence of response process heterogeneity in actual data and highlight the measurement problems we encounter when we assume all respondents conform to one response process.

In Chapter 4, I extend the investigation of item response process heterogeneity in noncognitive assessments with the use of process data, specifically response time data. This study empirically examines systematic differences in item-person interactions across fast and slow item responses in rating scale measurement, attending to both the intended-to-be-measured content trait and unintended-to-be-measured response styles. Through real data illustrations, the study demonstrates simultaneous relevance of both the content trait and response styles on response times and discusses implications for using response time information in noncognitive assessments.

While each of the three studies apply different psychometric modeling approaches to different types of assessments, the common goal is to develop a better understanding of item response process and behaviors in assessments. By extension, a primary objective of this dissertation is to provide demonstrations and frameworks for modeling item response processes which can ultimately lend insight into individual differences in measurement, contribute to diagnostic assessment, and enhance the overall validation of test items and tests. The following chapters will present the three studies in more detail followed by overall conclusions and future directions for research.

2 Noncompensatory MIRT for Passage-Based Tests [1]

2.1 Introduction

A passage-based test structure is commonly used for measuring cognitive proficiencies such as reading comprehension and analytical reasoning. Items in such tests are organized around passages, where a group of items related to a common passage is often referred to as a testlet (Wainer & Kiely, 1987). Passage-based tests tend to violate the local independence assumption of item response theory (IRT) when traditional unidimensional IRT models are fit, and, thus, can produce biased parameter estimates and standard errors (Bradlow et al., 1999; DeMars, 2006; Ip, 2000; Ip et al., 2013; Sireci et al., 1991; Thissen et al., 1989; Wainer & Thissen, 1996). Variants of testlet and bifactor models are often applied to address such violations, and their use in the context of various measurement applications have been extensively studied and developed (Bradlow et al., 1999; Cai et al., 2011; DeMars, 2006; Gibbons & Hedeker, 1992; Holzinger & Swineford, 1937; Li et al., 2006; Wainer et al., 2007). Importantly, there are different ways of addressing local dependence, not all of which entail multidimensionality (Ip, 2010; Yen, 1993). When viewed from a multidimensional IRT (MIRT) perspective, however, testlet and bifactor models incorporate general and specific factors, often viewed as the intended proficiency dimension (e.g., reading proficiency) and nuisance dimensions (e.g., content of passage), respectively. Such models fit within the category of *compensatory* MIRT models in the sense that the general and specific factors are assumed

[1] This chapter is an adapted version of Kim et al. (2022) published in *Psychometrika* and has been reproduced here with the permission of the copyright holder.

to interact additively. As a consequence, the influence of the general and specific factors function in compensating ways under these models (e.g., a high level of passage-specific knowledge offsets a lack of reading proficiency that is independent of the passage, and vice versa).

As an alternative, this chapter suggests it may be more appealing for many passage-based tests to assume a conjunctive form of interaction. Maris (1995) demonstrated the potential to estimate a model involving separable latent components that underlie response processes through a conjunctive Rasch model (basically a noncompensatory MIRT model), a modeling form also presented by Embretson (1984) and Whitely (1980). This study seeks to show how Maris (1995)' approach can be adapted to a passage-based test where passage-based items are affected by passage-related and passage-independent components that may interact in a noncompensatory way. For example, in a reading comprehension test, it is conceivable to separate underlying components related to a comprehension of the reading passage (i.e., passage-related) and processing of the specific item (i.e, passage-independent). Correct responses can be expected only if both components are successfully executed suggesting a noncompensatory interaction of the components. Specifically, to correctly respond to an item such as “What is the main idea of the passage?”, an examinee is required (1) to read and process the passage sufficiently well to understand the main idea, AND (2) to have a passage-independent understanding of what the concept of a main idea is. It is difficult to envision proficiencies underlying these components compensating for each other in the way reflected by a compensatory model.

Based on the idea above, I present a noncompensatory MIRT model for passage-based tests (Note that noncompensatory models are sometimes alternatively referred to as partially compensatory models, e.g., Reckase, 2009; like others, I choose the term “noncompensatory” to emphasize that compensation of proficiencies does not occur within any statistical component). A noncompensatory approach has several primary advantages. First, as described above, it arguably provides a more accurate account of the underlying cognitive processes/ components involved in responses to passage-based items. Beyond accounting for statistical dependence across items from the same passage, a noncompensatory representation can provide a psychologically meaningful representation of how respondents arrive at a correct response. Second, from a statistical standpoint, the model yields distinct difficulty parameters for each component (and associated dimension), which is not possible within compensatory MIRT models (Bolt & Lall, 2003). I suggest that the potential to attach separate difficulty parameters to components makes it easier to understand and communicate how the distinct components of a passage-based item separately contribute to the overall difficulty of the item, a feature that may be particularly appealing for test developers and practitioners. Third and finally, the noncompensatory model proposed in this study provides a way of addressing recent concerns over the tendency of the bifactor model to overfit data (Bonifay & Cai, 2017; Bonifay et al., 2017). As will be shown, a conjunctive Rasch representation consistently entails one fewer parameter per item compared to the bifactor model; I use simulation and a real data application to study the degree to which the simpler noncompensatory model can nevertheless provide a close comparable fit.

The remainder of this chapter is organized as follows. First, the subsequent two sections provide an overview of the bifactor model, its application to passage-based tests, and the proposed noncompensatory alternative model. I next present a fully Bayesian approach to the estimation of each model. I evaluate the parameter recovery of the noncompensatory model, as well as the capacity of the bifactor and noncompensatory models to fit datasets generated from each of the models while also attending to the overfitting feature of the bifactor model. Moreover, I demonstrate the way each model ultimately captures the relative influence of passage-independent and passage-related components on items. Lastly, I consider both models in a real data application and examine the similarities across models in how they quantify the relative contributions of passage-related components to items. I specifically demonstrate the value of the noncompensatory model in separating the sources of overall item difficulty by examining the relationship between item type and the relative difficulty of the two components.

2.2 A Bifactor MIRT Approach for Passage-Based Tests

Many standardized tests have a passage-based structure where a group of items is associated with a common stimulus. An example is a reading comprehension test where several items are based on a common reading passage. Passage-based tests often produce local item dependence when a traditional unidimensional IRT model is fit. Bifactor MIRT models (Cai et al., 2011; Gibbons & Hedeker, 1992; Holzinger & Swineford, 1937) and related testlet models (Wainer et al., 2007) have received much attention as a way of accounting for passage-related local dependence both due to their frequently good empirical fit as well as recent advances related to their estimation (e.g., Cai et al., 2011; Rijmen, 2009) that can make the models practical even in cases of high dimensionality. In this study, I mainly consider the bifactor model, as it can be viewed as a generalization of the testlet model (DeMars, 2006; Li et al., 2006).

Assume a passage-based test with K passages having a total of I items, where each item is associated with exactly one passage. The bifactor model introduces a single general factor θ_0 and K mutually orthogonal specific factors θ_k $(k = 1, \cdots, K)$. The probability of correct response to an item i (from passage k) is given by:

$$P(U_{i(k)} = 1 \mid \theta_0, \theta_k) = \frac{1}{1 + \exp\{-(a_{i0}\theta_0 + a_{ik}\theta_k + d_i)\}} \tag{2.1}$$

where $U_{i(k)} = 1$ denotes a correct response to item i (from passage k), a_{i0} and a_{ik} are, respectively, the item discrimination parameters for the general and specific factors, and d_i is an intercept parameter. Each item loads on the general factor and exactly one specific factor. As a result, the bifactor MIRT model effectively provides a decomposition of an item's shared latent item variance with respect to a dimension measured by all items (i.e., general factor θ_0) and a dimension specific to the passage on which the item is based (i.e., specific factor θ_k).

The bifactor model can be viewed as a compensatory MIRT model in that the general and specific factors interact such that an increase in one factor can compensate for a deficiency in the other, with the degree of compensation affected by the magnitudes of the discrimination parameters. This compensatory nature of the model can be clearly seen from the linear combination of the factors in Equation (2.1). The model attaches separate discrimination parameters to each factor. Naturally, the relative size of discrimination parameters across dimensions (i.e., a_{ik}/a_{i0}) helps us understand the magnitude of the specific factor effect relative to the general factor effect for each item. Note the bifactor model reduces to the two-parameter version of a testlet model if this ratio for items is fixed to a common value within passages (DeMars, 2006; Li et al., 2006). Due to its compensatory nature, the bifactor model cannot uniquely identify the item difficulty for each dimension. Instead, each item can be viewed as having a single multidimensional difficulty parameter (MDIFF), defined as $-d/\sqrt{\mathbf{a}'\mathbf{a}}$ where d is the intercept and **a** is the vector of discrimination parameters for the item that effectively scale the intercept (Reckase, 2009).

One empirical observation from the application of bifactor models to passage-based tests is the tendency for certain item types to demonstrate a higher degree of passage-related dependence than others (Li et al., 2006). This result is apparent from systematically higher discrimination estimates seen on the specific factors relative to the general factor (i.e., large a_{ik}/a_{i0}), indicating the items are more "influenced" by the specific factor. Examples of such item types in the reading comprehension test studied by Li et al. (2006) include Inference or Main Idea items, which usually require a deeper level of passage-related processing. By contrast, item types such as Vocabulary in Context generally required a lower level of passage-related processing. It is important to note that these results are derived from the patterns of discrimination estimates for the general and specific factors in the bifactor model. Thus, the results speak less directly to the relative difficulties of the two processes/components (i.e., passage-related and -independent) on solving the item. I, therefore, consider a noncompensatory approach as attractive in part due to its potential to separately estimate the difficulty of each component.

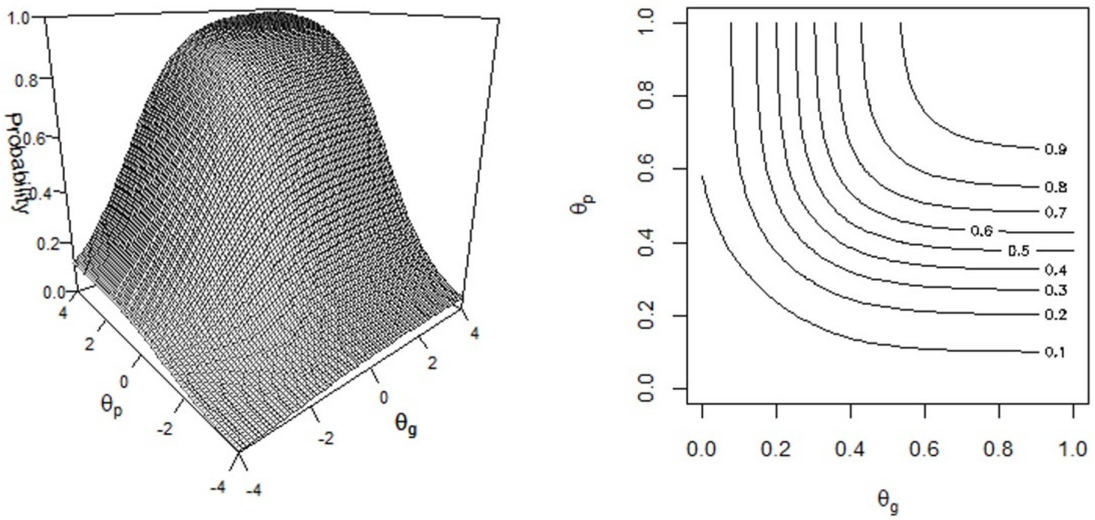
2.3 Noncompensatory Model for Passage-Based Tests

A Passage-Based Noncompensatory Model (PB-NM) is proposed below as an alternative to the bifactor model. I assume the presence of two underlying components to solving each passage-based item, a first component connected to the processing of the passage (and the extraction of all needed information from the passage for the particular item) and a second component related to passage-independent requirements of the item. The two passage-related and passage-independent components can take place in any order and need not be viewed as sequential. For a given item i (from passage k), a passage-based ability θ_{pk} $(k = 1, \cdots, K)$ is assumed to underlie the first component while a general ability θ_g is assumed to underlie the second passage-independent component. As items from the same passage likely entail the extraction of different information, the first passage-related component is modeled as distinct across items (even though based on the same θ_{pk}). As a result, the probability of correct response to item i (from passage k) can be written as:

$$P(U_{i(k)} = 1 \mid \theta_g, \theta_{pk}) = P(X_{ip} = 1 \mid \theta_{pk})P(X_{ig} = 1 \mid \theta_g) \tag{2.2}$$

where $X_{ip} = 1$ and $X_{ig} = 1$ denote the successful execution of passage-related and -independent components, respectively, for item i. The product of the two terms implies that both components need to be successfully performed to produce a correct response, and the overall probability of correct response on the item cannot exceed the minimum of the two terms, reflecting a noncompensatory interaction of the two components. A graphical representation to help understand the model (Ackerman, 1996) is shown in Figure 2.1. Due to the noncompensatory nature of the model, the item surface plot of the PB-NM has curved (as opposed to linear) equiprobable contour lines.

Figure 2.1
Illustration of a Surface Plot (Left) and a Contour Plot (Right) for Correct Response Probability of an Item Under the Passage-Based Noncompensatory Model (PB-NM)



Note. θ_g = Passage-independent ability; θ_p = Passage-related ability.

As for the bifactor model, each examinee has a single general proficiency θ_g and K passage-related proficiencies θ_{pk}. It is assumed that the ability related to the passage, θ_{pk}, stems from the general proficiency, θ_g, with a passage-specific shift (as might be due, for example, to the content of the passage involved). The noncompensatory structure in Equation (2.2) can use different parametric forms for the two components. In this study, I use a Rasch model to represent each component. The PB-NM, therefore, can be specified as:

$$P(U_{i(k)} = 1 \mid \theta_g, \theta_{pk}) = \frac{1}{1 + \exp\{-(\theta_{pk} - b_{ip})\}} \times \frac{1}{1 + \exp\{-(\theta_g - b_{ig})\}} \tag{2.3}$$

where b_{ip} and b_{ig} are, respectively, item difficulty parameters for passage-related and passage-independent components of item i. Note that beyond adapting the functional form of the bifactor approach, the PB-NM simultaneously seeks to reduce its parameterization, following the findings of Bonifay and Cai (2017) and Bonifay et al. (2017). Thus, from another perspective, the presented model examines whether an adequate account of the psychometric functioning of passage-based items can be achieved using two (as opposed to three) parameters per item.

As implied above, an appealing aspect of the noncompensatory representation in comparison to the bifactor representation is the potential to characterize the component processes in terms of the difficulties of the two components. The notion of using difficulty differences across dimensions to understand measurement of the dimensions has also been considered in relation to the multidimensional latent trait model (MLTM; Whitely, 1980) where the nature of items can be explained by attending to differences in the relative difficulties of items across dimensions. As observed for a different noncompensatory model in Bolt and Lall (2003), I anticipate a close empirical relationship between the difficulty estimates of the PB-NM and discrimination estimates of the bifactor model. It is specifically expected that the relative size of the item parameters across two dimensions under the PB-NM and the bifactor model may similarly represent the relative influence of passage-related and passage-independent components on item scores. As the difficulty representation under the PB-NM can be viewed as more immediately interpretable by applied practitioners, a high correspondence between the parameters under the two models may also support the noncompensatory approach. In addition, with two parameters (as opposed to three) per item, the PB-NM is expected to mitigate some of the concerns of overfit under the bifactor model (Bonifay & Cai, 2017; Bonifay et al., 2017).

As an additional note, I clarify that the PB-NM is different from the noncompensatory testlet model presented by Jiao et al. (2017) that also considers a noncompensatory interaction between proficiencies. The noncompensatory testlet model (Jiao et al., 2017) deals with the local dependence across items by assuming noncompensatory interactions across passage proficiencies. The PB-NM, however, attends to a noncompensatory interaction between passage-related and passage-independent components in solving passage-based items.

2.4 Bayesian Estimation Algorithm

The PB-NM and the bifactor model are fitted using a fully Bayesian approach through a Markov Chain Monte Carlo (MCMC) algorithm. For purposes of estimation, proficiency parameter θ_{pk} in the PB-NM is re-parameterized as $\theta_g + \eta_{pk}$, where η_{pk} is normalized within subject. Specifically, η'_{pk} is sampled from $Normal(\theta_{jg}, \sigma^2_{pk})$ and $\eta_{pk} = \eta'_{pk} - \frac{1}{K}\sum_{k=1}^{K} \eta'_{pk}$. Thus, θ_g and θ_{pk} will tend to correlate positively across subjects with η_{pk} representing a shift in proficiency reflecting a passage effect. The prior distributions used for parameters of the PB-NM are: $\theta_g \sim Normal(0,\ 1)$, $\eta'_{pk} \sim Normal(\theta_g, \sigma^2_{pk})$, $b_{ig} \sim Normal(0, 1)$, $b_{ip} \sim Normal(0, 1)$, and $1/\sigma^2_{pk} \sim Gamma(1, 1)$.

A similar parameterization is applied to the proficiencies of the bifactor model to make the algorithms for the two models similar. I define $\theta_k = \theta'_k - \frac{1}{K}\sum_{k=1}^{K} \theta'_k$ and sample θ'_k from $Normal(\theta_0, \sigma^2_k)$ so as to enforce orthogonality between the general and specific factors. Note that the specific factor θ_k is, thus, uncorrelated with the general factor θ_0 and is distinct from the passage-related factor θ_{pk} in the PB-NM, rather having a similarity to the passage-specific shift parameter in the PB-NM. Prior distributions for the bifactor model were specified as: $\theta_0 \sim Normal(0, 1)$, $\theta'_k \sim Normal(\theta_0, \sigma_k^2)$, $a_{i0} \sim logNormal(0, 0.25)$, $a_{ik} \sim logNormal(0, 0.25)$, $d_i \sim Normal(0, 1)$, and $1/\sigma_k^2 \sim Gamma(1, 1)$.

For both the PB-NM and the bifactor model, I simulated 10 parallel chains where initial values were randomly generated from the prior distributions of the parameters. For each simulated chain, the first 20,100 iterations are ignored (100 as adaptive phase iterations and the next 20,000 as burn-in iterations) and the samples from a subsequent 20,000 iterations are monitored to construct posterior distributions. A thinning interval of 10 iterations was used, so 2,000 iterations from the respective chain (2,000×10=20,000 iterations in total) were used to define the posterior distributions of the parameters. These analyses were implemented in JAGS (Just Another Gibbs Sampler) 4.3.0 (Plummer, 2017) using R (R Core Team, 2020) through *jagsUI* (Kellner, 2019). In addition to visual inspection of the sampling chains, the Gelman-Rubin diagnostic R^2 (Gelman & Rubin, 1992) was used as a convergence criterion.

2.5 Simulation Study

A small scale simulation was conducted to examine both (1) parameter recovery for the PB-NM, and (2) the empirical relationship between the PB-NM and the bifactor model parameter estimates when fitted to common datasets. Through the simulation, I seek to show that the PB-NM is in fact estimable for passage-based tests under the types of conditions commonly observed when fitting bifactor models. Similar to Bolt and Lall (2003), I also seek to show that proportional differences in general and specific factor discriminations under the bifactor model similarly emerge as differences in difficulties under PB-NM, and vice versa.

Data were generated from both the PB-NM and the bifactor model following Equations (2.3) and (2.1). For the PB-NM, the general proficiency parameter θ_g was generated from $Normal(0, 1)$ and the passage-related proficiency θ_{pk} was defined as $\theta_g + \eta_{pk}$, where $\eta_{pk} = \eta'_{pk} - \frac{1}{K}\sum_{k=1}^{K} \eta'_{pk}$ and η'_{pk} were generated from $Normal(\theta_g, \sigma^2_{pk})$. The variance parameter σ^2_{pk} was generated from $Gamma(1, 1)$ and the b_{ig} and b_{ip} for both components were independently generated from $Uniform(-3, 3)$. The generating distributions for bifactor parameters were $[\theta_0, \theta_1, \cdots, \theta_K] \sim MVN(0, I)$, $a_{i0} \sim Uniform(0.5, 2.5)$, $a_{ik} \sim Uniform(0.5, 2.5)$, and $d_i \sim Uniform(-2, 2)$. These parameter distributions were chosen to produce varying combinations of item parameters across the two dimensions. I believe such distributions can, to some extent, reflect the varying passage-based item types often used for reading comprehension tests, and may or may not be applicable to other types of passage-based tests not having such characteristics. Besides the two generating models, my simulation design considered factors of sample size ($N = 1000, 2500$) and number of items per passage ($N_{i/p} = 4, 6, 9$), holding the total number of items to 36. All three factors were crossed, yielding 12 conditions. Ten datasets were generated for each simulation condition.

I fitted both the PB-NM and bifactor model to each dataset. One of the ten datasets generated under the PB-NM for the condition having 4 items per passage and 2,500 examinees did not converge when the PB-NM was fitted, seemingly due to a dimension switching problem (Bolt & Lall, 2003). Such an observation suggests that non-convergence can possibly happen if the two proficiencies across the dimensions are highly correlated or are difficult to distinguish; in my case, the small variance of passage-specific shift parameter corresponding to the passage-related factor as well as the small number of items per passage (both of which make it difficult to detect passage effects) likely produced the problem. Some possible solutions to resolve the convergence issue in practice could include simplifying the PB-NM model (e.g., fix the variance of passage-related factor to be identical across passages) or imposing ordinal constraints on the item difficulty parameters across components to prevent dimension switching. In this study, the non-converging case was excluded when deriving the results as it was the only replication that produced a problem. I focused on three primary outcomes from these analyses: item parameter recovery for the PB-NM (when data were generated from the PB-NM), the relative fit of each model to each dataset, and the relationships between the resulting difficulty and discrimination estimates across models.

2.5.1 Item Parameter Recovery for the PB-NM

The recovery of item difficulty parameters under the PB-NM was examined in two ways. Because of the interest in interpreting the relative difficulty of the components, I first examined the recovery of the within-item difference in difficulty parameters across dimensions. Specifically, the correlation of the estimates and parameters for the difference in difficulties of the two components (i.e., $b_{\text{diff},i} = b_p - b_g$) was obtained under each simulation condition. I also inspected the recovery of individual item difficulty parameters by deriving bias and the root mean square error (RMSE) indices defined as

$$\text{Bias}_i = \frac{1}{R}\sum_{r=1}^{R}\hat{b}_{icr} - b_{ic} \tag{2.4}$$

and

$$\text{RMSE}_{\,i} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{b}_{icr} - b_{ic})^2} \tag{2.5}$$

where $\hat{b}_{icr}$ denotes item i's difficulty estimate for component c (i.e., p or g) from the rth dataset, and R denotes the total number of replications in each cell (10 in this study).

2.5.2 Comparison of the PB-NM and the Bifactor Model

I compared the fit of the two models under each data generating condition using the deviance information criteria (DIC). The DIC penalizes model complexity by adding an effective number of parameters (p_D) term to the mean deviance (i.e., $DIC = \overline{D(\theta)} + p_D$). I further examined the degree of correspondence seen between the parameter estimates of the two models in their characterization of the relative influence of passage-related and passage-independent components on item performance. I specifically examined the relationship between the within-item difference in difficulty estimates ($\widehat{b}_{\text{diff},i}$) under the PB-NM and the within-item relative discrimination estimates, calculated as the log of the ratio of specific to general factor discrimination estimates, under the bifactor model (i.e., $\widehat{a}_{\text{ratio},i} = log(\hat{a}_{ik}/\hat{a}_{i0})$), by deriving the correlation of the two indices across items.

2.6 Simulation Results

The findings of the simulation study are organized into subsections focusing on the parameter recovery of PB-NM, model fit comparisons of the PB-NM against the bifactor model, and the relationship between the within-item parameter differences for the two models.

2.6.1 Item Parameter Recovery

Table 2.1 and Figure 2.2 illustrate parameter recovery results for the PB-NM. The results make apparent a couple of issues. First, as expected, item parameters are recovered better when there is a larger number of respondents. Recovery is also consistently better when there is a relatively large number of items per passage, which in the current simulation design also directly corresponds to a smaller overall number of dimensions. This result is not surprising given that with a larger number of items per passage, the passage-related dimensions become better estimated and, hence, the ability to distinguish difficulty levels across items within those dimensions is seemingly better. The number of respondents, however, seems to be the more critical factor.

It is also observed from Figure 2.2 that the bias likely occurs because the parameter estimates are shrunk to the prior mean of 0. Moreover, the bias appears greater for easier components, particularly for those items that are difficult on the other component (see Figure A1 in Appendix A). This result is consistent with DeMars (2016)' findings that, when factors are highly correlated, the direction and magnitude of the bias in difficulty parameters are influenced by the relative size of difficulties across dimensions. As in Figure A1, DeMars (2016)' results specifically show that the easier component is positively biased while the more difficult component is negatively biased, and the bias increases as the discrepancy between difficulties across dimensions gets larger, especially for the easier component. Given the non-compensatory interaction of the components as well as the positively correlated trait factors, I suspect this may be due to the difficult component reducing the information available for

Table 2.1
Recovery of Item Difficulty Parameters for Passage-Based Noncompensatory Model (PB-NM)

N	Criterion		$N_{i/p}$		
			4	6	9
1,000	$corr(b_{\text{diff},i}, \widehat{b_{\text{diff},i}})$		0.663	0.885	0.891
	Absolute bias	$\hat{b}_p$	0.702	0.549	0.479
		$\hat{b}_g$	0.720	0.537	0.471
	RMSE	$\hat{b}_p$	0.799	0.662	0.618
		$\hat{b}_g$	0.797	0.628	0.593
2,500	$corr(b_{\text{diff},i}, \widehat{b_{\text{diff},i}})$		0.910	0.974	0.969
	Absolute bias	$\hat{b}_p$	0.336	0.287	0.165
		$\hat{b}_g$	0.340	0.225	0.148
	RMSE	$\hat{b}_p$	0.473	0.398	0.310
		$\hat{b}_g$	0.451	0.304	0.266

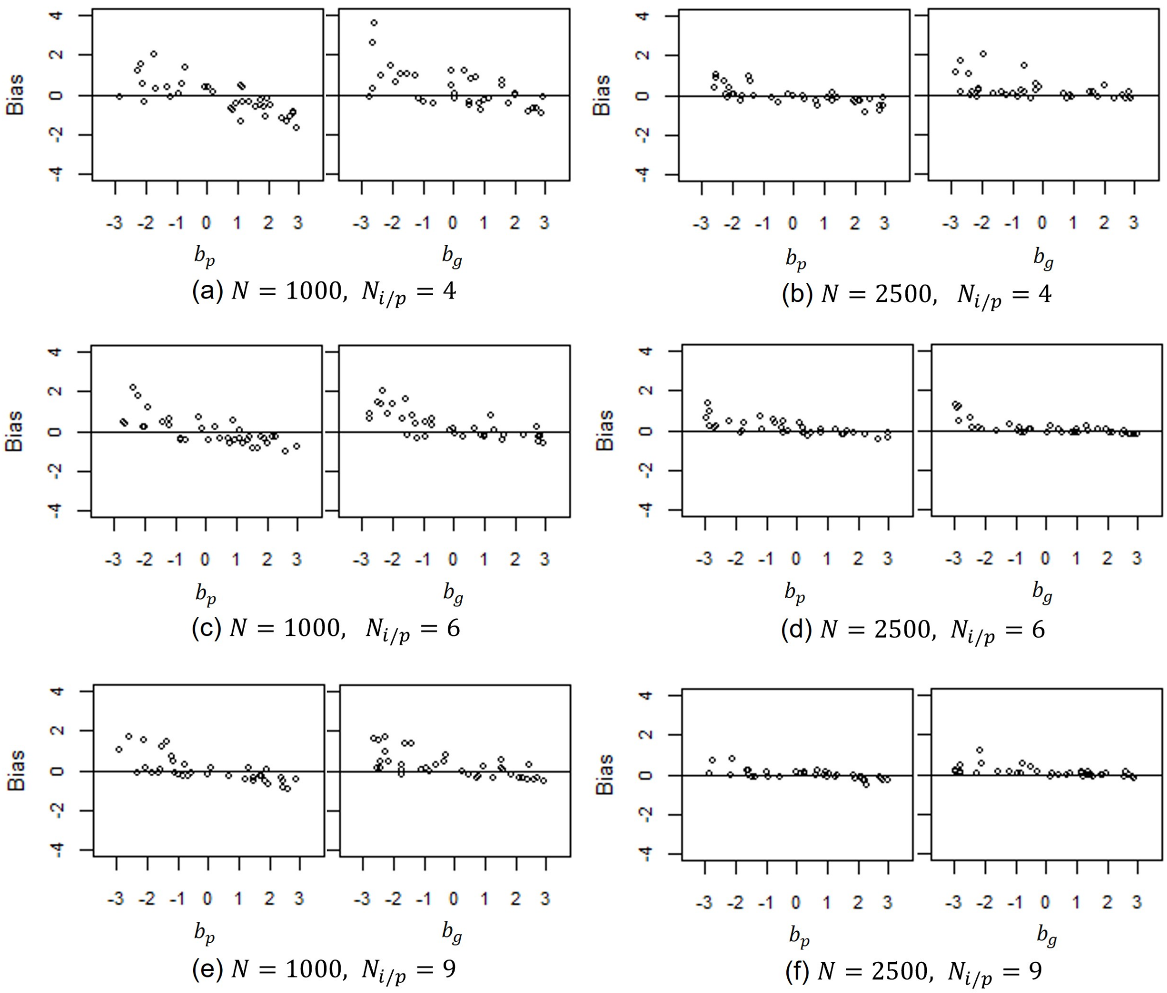
Note. The correlation coefficients are averaged across replications. The absolute bias and RMSE are averaged across items. N = The number of respondents; $N_{i/p}$ = The number of items per passage; b_p = Passage-related difficulty; b_g = Passage-independent difficulty; $b_{\text{diff},i}$ = Difference in difficulty parameters across dimensions ($b_p - b_g$); $\widehat{b_{\text{diff},i}}$ = Difference in difficulty estimates across dimensions ($\hat{b}_p - \hat{b}_g$); RMSE = Root mean square errors.

estimating the difficulty parameter for the "easy" component, a result also discussed by Maris (1995).

Furthermore, it appears that the capability to interpret the difference in item difficulties under the PB-NM, as conveyed by $corr(b_{\text{diff},i}, \widehat{b_{\text{diff},i}})$, is relatively well-achieved, especially when the sample size is large. In five of the six conditions considered, the correlation exceeds .88. The one exception concerns conditions with a smaller number of items per passage (4) and smaller sample size (1000), where the correlation drops to .66.

Figure 2.2

The Bias of Estimates of Difficulty Parameters Under the Passage-Based Noncompensatory Model (PB-NM), Fitted to Data for Each Generating Condition



(a) $N = 1000,\ N_{i/p} = 4$

(b) $N = 2500,\ N_{i/p} = 4$

(c) $N = 1000,\ N_{i/p} = 6$

(d) $N = 2500,\ N_{i/p} = 6$

(e) $N = 1000,\ N_{i/p} = 9$

(f) $N = 2500,\ N_{i/p} = 9$

Note. N = The number of respondents; $N_{i/p}$ = The number of items per passage; b_p = Passage-related difficulty; b_g = Passage-independent difficulty.

2.6.2 Model Fit Comparison with Bifactor Model

Table 2.2 reports DIC comparison results for the PB-NM and the bifactor model. As can be seen in the table, the bifactor model consistently has a larger p_D, indicating a greater model complexity, than the PB-NM. This observation is consistent with the overfitting tendencies of the model (Bonifay & Cai, 2017; Bonifay et al., 2017). The p_D consistently shows larger discrepancies between models as the number of items per passage reduces (and thus, under this simulation design, the larger number of dimensions in the overall model). As anticipated, in terms of both mean deviance and DIC values, the PB-NM always fit better for the data generated under the PB-NM. For data generated under the bifactor model, the bifactor model consistently fit better in terms of mean deviance, but only with respect to DIC when there was a sufficiently large number of items per passage (i.e., larger than 4). When there was a small number of items per passage ($=$ 4), the PB-NM had smaller DIC. This implies that, with a small number of items per passage and, thus, more overall dimensions, the cost of complexity may exceed the value of increased fit under the bifactor model even when the bifactor model is the generating model. It would appear that, under such conditions, the PB-NM can effectively capture the nature of the passage-related dependence while reducing the overfit that occurs under the bifactor model.

2.6.3 Comparison of Relative Item Parameter Estimates

Despite their different statistical representations, the PB-NM and the bifactor model are suspected to provide similar information regarding the relative effects of passage-related and passage-independent factors on item responses. Table 2.3 reports the mean correlations observed between the difference in difficulty estimates ($\widehat{b}_{\text{diff},i}$) under the PB-NM and the relative discrimination estimates ($\widehat{a}_{\text{ratio},i}$) under the bifactor model. As anticipated, high correlations were observed. While a consistent effect of the manipulating factors is not observed (note that by holding the overall number of items constant, increasing the number of items per passage

Table 2.2

Averaged Mean Deviance and Deviance Information Criterion (DIC) Across 10 Replications, for Passage-Based Noncompensatory Model (PB-NM) and Bifactor Model, Fitted to Data for Each Generating Condition

Condition		Fitted model	Generating model					
N	$N_{i/p}$		PB-NM			Bifactor model		
			p_D	$\overline{D(\theta)}$	DIC	p_D	$\overline{D(\theta)}$	DIC
1,000	4	PB-NM	9518.3	23936.9	**33455.2**	12669.5	31140.2	**43809.6**
		Bifactor	11769.9	23967.7	35737.6	14969.8	29934.0	44903.8
	6	PB-NM	6722.9	26492.7	**33215.6**	7462.0	29963.8	37425.8
		Bifactor	7160.9	26618.2	33779.1	7726.9	28726.4	**36453.3**
	9	PB-NM	3978.8	26593.4	**30572.2**	4825.4	30881.4	35706.7
		Bifactor	4438.9	26662.8	31101.6	4899.3	29851.7	**34751.0**
2,500	4	PB-NM	27501.9	64361.6	**91863.5**	31137.6	75952.2	**107089.8**
		Bifactor	43572.0	64817.6	108389.7	37033.7	72656.9	109690.6
	6	PB-NM	18774.6	65855.9	**84630.5**	18330.2	82153.0	100493
		Bifactor	20704.9	66553.5	87258.4	19638.6	79733.0	**99371.5**
	9	PB-NM	11254.4	55688.7	**66943.1**	11304.9	84208.6	95513.5
		Bifactor	11644.4	56150.2	67794.6	11899.7	81469.6	**93369.4**

Note. The smaller value of DIC is emboldened. N = The number of respondents; $N_{i/p}$ = The number of items per passage; p_D = The effective number of parameters; $\overline{D(\theta)}$ = Mean deviance.

decreases the number of passages, and vice versa), the correlations are larger than 0.67 under all conditions. There also appear to be slightly higher correlations between estimates when the data are generated from the PB-NM than the bifactor model, a result that is likely due to the additional item parameter per item included under the bifactor model. Overall, however, the consistently high correlations imply that difficulty parameters of the PB-NM and discrimination parameters of the bifactor model provide similar information regarding the relative size of passage-related effects in the items.

Table 2.3

Correlation Between the Difference in Difficulty Estimates ($\widehat{b}_{diff,i}$) Under the Passage-Based Noncompensatory Model (PB-NM) and the Relative Discrimination Estimates ($\widehat{a}_{ratio,i}$) Under the Bifactor Model, Averaged Across 10 Replicated Datasets for Each Data Generating Condition

$\boldsymbol{N}$	$\boldsymbol{N_{i/p}}$	**Generating model**	
		PB-NM	Bifactor
1,000	4	0.852	0.882
	6	0.838	0.742
	9	0.876	0.808
2,500	4	0.872	0.668
	6	0.930	0.782
	9	0.950	0.713

Note. N = The number of respondents; $N_{i/p}$ = The number of items per passage.

2.7 Real Data Illustration

2.7.1 Methods

The real data application considers a reading comprehension test administered as part of a university placement test at a Midwestern university. The dataset includes 36 items administered to 22,510 students coming from a quite homogeneous population with not much ethnic diversity. The 36 items are nested in 8 reading passages; the number of items corresponding to passages 1 through 8 are respectively 4, 5, 4, 5, 5, 4, 4, and 5. In this study, I used responses from 2,500 randomly sampled students to reduce computational burden. The reading comprehension test consists of items categorized by several types, including Literal Comprehension (LC), Main Idea (MI), Inference (INF), Vocabulary in Context (VOC), Rhetorical Strategy (RS), Development (DEV), Tone (TON), and Logical Reasoning (LR). Example questions for each item type are given in Appendix B.

I evaluated the presence of local dependence among items within passages by examining the Jackknife Slope Index (JSI; Edwards et al., 2018) under a unidimensional two-parameter logistic (2PL) model. The JSI is a recently proposed tool for diagnosing the potential presence of local dependence between pairs of items. Edwards et al. (2018) has reported that it has high power and reasonably low Type I error rates in comparison to other indices such as Q_3 (Yen, 1984) and G^2 (W. H. Chen & Thissen, 1997). The JSI values were produced using *mirt* (Chalmers, 2012) in R (R Core Team, 2020) and 15 item pairs were flagged as locally dependent in the reading comprehension test data. Most of the flagged item pairs (12 out of 15) occurred within passages suggesting a presence of local item dependence within passages. I also compared the model fit across unidimensional models (i.e., Rasch and 2PL models) and the bifactor model to confirm the existence of local item dependence. As shown in Table A1, the superior model fit of the bifactor model supported the presence of meaningful local dependence among items.

After confirming the presence of local dependence in the data, I fitted each of (1) the PB-NM and (2) the bifactor model using the same fully Bayesian estimation approach described above. Then I compared the model fit using the deviance information criterion (DIC) values and, as in the simulation, also compared models in terms of their parameter estimates. In addition, I tried to confirm whether the component difficulty estimates conform to the theoretical expectations in terms of the different item types. It was anticipated that some item types would consistently have higher or lower difficulties on the passage-related component relative to that of the passage-independent component. For example, item types such as Main Idea should show higher difficulty on the passage-related component due to the clear need for passage comprehension, whereas Vocabulary items will likely show greater relative difficulty on a passage-independent component due to the reduced need for comprehension of the passage.

2.7.2 Results

Table 2.4 provides empirical comparative fit results for the PB-NM and the bifactor model. While the bifactor model shows a slightly lower mean deviance, as in the simulation results, the difference is more than offset by its complexity, and the likelihood of overfitting the data is apparent. The PB-NM, thus, appears at least as good as the bifactor model in explaining these data, a result consistent with what was seen in the simulation results. A preference for the PB-NM may particularly hold when the number of items per passage is small, where the overfit of the bifactor model is most noticeable.

Table 2.4
A Model Fit Comparison for Passage-Based Noncompensatory Model (PB-NM) and Bifactor Model, for Reading Comprehension Test Data

Model	p_D	$\overline{D(\theta)}$	DIC
PB-NM	24475.1	87488.1	**111963.2**
Bifactor Model	37756.5	86031.9	124414.6

Note. The smaller value of Deviance Information Criterion (DIC) is emboldened. p_D = The effective number of parameters; $\overline{D(\theta)}$ = Mean deviance.

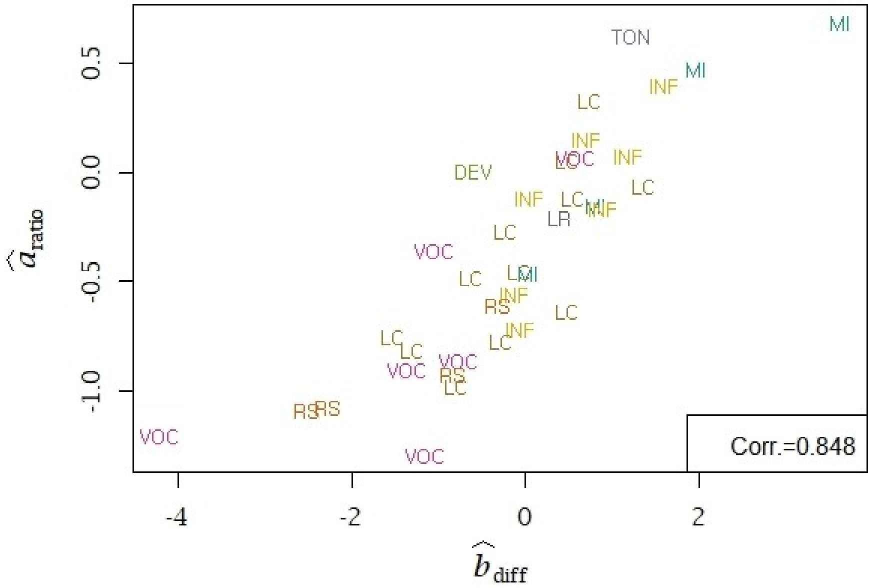
As in the simulation, a close relationship between item parameter estimates across the PB-NM and the bifactor model was also observed. The difference in difficulty estimates ($\widehat{b}_{\text{diff},i}$) under the PB-NM and the relative discrimination estimates ($\widehat{a}_{\text{ratio},i}$) derived from the bifactor model were highly correlated (correlation = 0.848). Thus, items with a high relative bifactor discrimination on the specific factor (i.e., high $\widehat{a}_{\text{ratio},i}$) tended to have a high difficulty on the passage-related factor compared to the passage-independent factor (i.e., high $\widehat{b}_{\text{diff},i}$) under the PB-NM. This indicates that, irrespective of the overall item difficulty, the item having a relatively difficult passage-related component (in comparison to the passage-independent component) is likely to be an item with a large passage effect, and vice versa.

The results also suggested an association between item type and the relative size of within-item parameter estimates under both models. Figure 2.3 displays a scatterplot of the relative discriminations ($\widehat{a}_{\text{ratio},i}$) under the bifactor model and difficulty differences ($\widehat{b}_{\text{diff},i}$) under the PB-NM with item type labels. Similar to the results under the bifactor model in Li et al. (2006), item types such as Vocabulary and Rhetorical Strategy, not expected to have a strong passage-related influence, were mostly located in the lower left part of the plot. This indicates that these items have relatively small passage-specific discriminations and easy passage-related components under both the bifactor model and PB-NM, respectively. Alternatively, item types such as Main Idea and Inference which are expected to show stronger passage-related influence, tend to be found in the upper right part of the plot. At the same time, there are individual cases where the Vocabulary and Inference items, respectively, have somewhat more difficult and easier passage-related components relative to the passage-independent component. Such instances likely reflect cases where Vocabulary items may ask about a word whose meaning is more dependent on context, or in the case of inference an item whose passage is easier to understand.

The analysis of variance (ANOVA) results in Table 2.5 statistically confirms the association by showing that the means of the relative size of item parameters do differ across

Figure 2.3

A Plot of Difference in Difficulty Estimates ($\widehat{b}_{diff,i}$) Under the Passage-Based Noncompensatory Model (PB-NM) Against Relative Discrimination Estimates ($\widehat{a}_{ratio,i}$) Under the Bifactor Model with Item Type Labels, for Reading Comprehension Test Data



Note. LC = Literal comprehension; MI = Main idea; INF = Inference; VOC = Vocabulary in context; RS = Rhetorical strategy; DEV = Development; TON = Tone; LR = Logical reasoning.

item types under both models. Note that I did not include items types of Deviation, Logical Reasoning, and Tone in the analyses because there was only one item measured for each of those item types. In Figure 2.4, boxplots of the $\widehat{b}_{\text{diff},i}$ and $\widehat{a}_{\text{ratio},i}$ are shown for each item type under the two models. It is consistently observed from the figure that item types such as Main Idea and Inference have high $\widehat{b}_{\text{diff},i}$ and $\widehat{a}_{\text{ratio},i}$ under the two models, while item types such as Vocabulary and Rhetorical Strategy have low values. The mean values (shown as red X) also appear to have a very similar pattern across the two models, again illustrating that the difference in difficulties under the PB-NM and the relative discriminations under the bifactor model distinguish the items in similar ways.

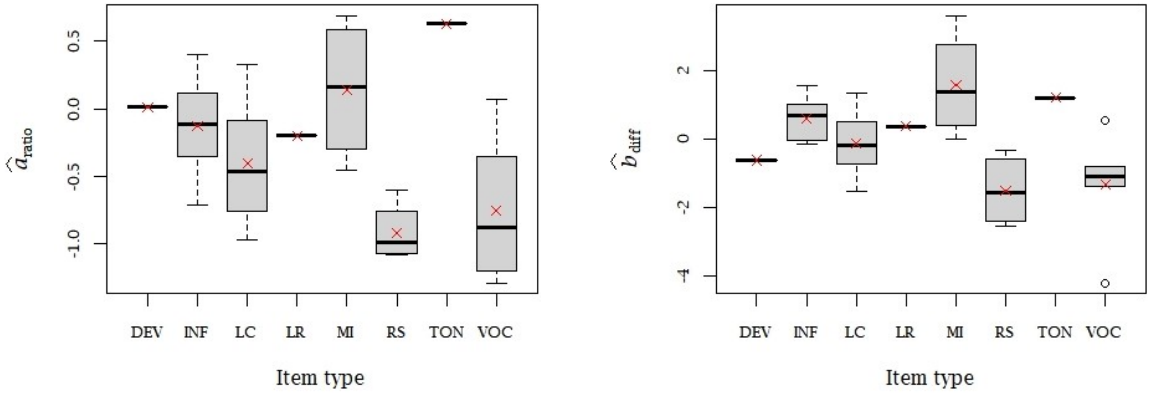
Table 2.5

The Analysis of Variance (ANOVA) Results for Testing Differences in the Means of Difference in Difficulty Estimates ($\widehat{b}_{diff,i}$) and the Means of Relative Discrimination Estimates ($\widehat{a}_{ratio,i}$) Across Item Types, Each Under the Passage-Based Noncompensatory Model (PB-NM) and the Bifactor Model

		SS	***df***	***MS***	***F***	***p***
Bifactor	Between	3.52	4	0.88	4.85	0.0043
	Within	5.09	28	0.18		
PB-NM	Between	31.58	4	7.90	6.51	0.0008
	Within	33.96	28	1.21		

Figure 2.4

Boxplots of Relative Discrimination Estimates ($\widehat{a}_{ratio,i}$) Under the Bifactor Model and Difference in Difficulty Estimates ($\widehat{b}_{diff,i}$) Under the Passage-Based Noncompensatory Model (PB-NM), for Each Item Type, for Reading Comprehension Test Data

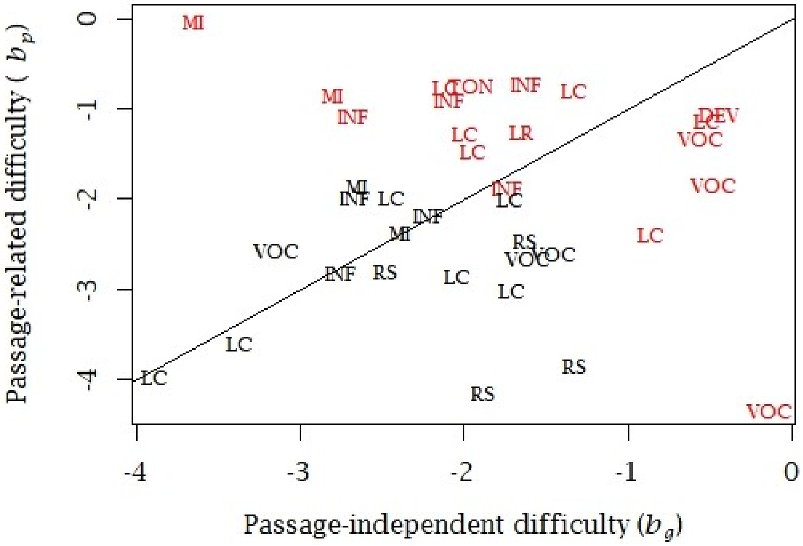


Note. Red cross indicates the mean value. LC = Literal comprehension; MI = Main idea; INF = Inference; VOC = Vocabulary in context; RS = Rhetorical strategy; DEV = Development; TON = Tone; LR = Logical reasoning.

Figure 2.5 further illustrates how the PB-NM isolates the sources of item difficulty and their relationship to item type. Here the items are identified by item type. Items that are difficult in a classical sense and have a p-value (i.e., proportion correct) smaller than the median p-value are denoted in red. This plot shows the ability of the PB-NM to differentiate the component difficulties of each item, and from the plot, we can identify which component is making the item more difficult, particularly for items in red. We can see, for example, that a Vocabulary item at the bottom right was difficult because of a difficult passage-independent component, in spite of an easy passage-related component. By contrast, a Main Idea item on the top left was difficult due to a difficult passage-related component despite having an easy passage-independent component. Such a separation of the difficulties conforms to our expectations given the greater or less passage-processing demands associated with certain item types. Specifically, item types such as Main Idea and Inference are generally located above the line, meaning that passage-processing for item solving is relatively more demanding, and item types such as Vocabulary are below the line, indicating that passage-processing is relatively less demanding. Thus, for the former item types, the overall item difficulty is often largely determined by the passage-related difficulty but not much by the passage-independent difficulty (especially for difficult items), and vice versa for the latter item types (see Figure A2).

Figure 2.5

A Scatter Plot of Passage-Related Difficulty and Passage-Independent Difficulty Under the Passage-Based Noncompensatory Model (PB-NM) with Item Type Labels, Difficult Items with Low p-Values (i.e., Smaller than the Median) Denoted in Red



Note. p-Value = Classical item difficulty (proportion-correct). LC = Literal comprehension; MI = Main idea; INF = Inference; VOC = Vocabulary in context; RS = Rhetorical strategy; DEV = Development; TON = Tone; LR = Logical reasoning.

In summary, the PB-NM seemed to return nearly equivalent fit for the reading comprehension passage-based test data with less complexity than the bifactor model. The relative size of item estimates derived from the PB-NM and the bifactor model were found to be highly correlated, and these values were closely related to item types in theoretically expected ways. Specifically, while the overall difficulty of the item is reflected by overall higher b parameters in the PB-NM model, the relative difference in the b parameters across components corresponds to the relative discriminations of the items on the general versus specific proficiency traits. Thus, it would appear that the PB-NM preserves useful information about the relative effects of passage-related and passage-independent proficiencies on item performance, and in a way (e.g., in terms of item difficulty) that is intuitively more interpretable than the bifactor approach.

2.8 Discussion and Conclusion

Local item dependence often arises when items are clustered around passages within a test. Among several approaches to modeling the local dependence (Cai et al., 2011; Gibbons & Hedeker, 1992; Hoskens & De Boeck, 1997; Ip, 2002; Thissen et al., 1989; Wainer et al., 2007), the bifactor model has been found to be a useful strategy. One of the appealing features of the bifactor model is the potential to examine the extent to which individual items are sensitive to specific versus general factors (Li et al., 2006). However, the need to interpret such effects in relation to discrimination parameters, as opposed to difficulty parameters under a componential IRT approach, may make the bifactor model less easily interpreted by practitioners. Arguably, a noncompensatory perspective allows practitioners to better appreciate the likely presence of two correlated components in solving passage-based items, as virtually all items can be understood as requiring some level of passage comprehension and some level of passage-independent item interpretation. Moreover, from a statistical perspective, the bifactor approach tends to overfit the data (Bonifay & Cai, 2017; Bonifay et al., 2017).

Another potentially questionable feature of the bifactor approach follows from its statistical representation as a compensatory MIRT model. Although the ability to empirically distinguish between compensatory and noncompensatory interactions is difficult if not impossible, there may be good reason to speculate that noncompensatory interactions provide a closer match to the components underlying passage-based items. Intuition would suggest that if the level of general reading proficiency is not sufficient, the potential for high passage-related proficiencies being able to completely offset the deficiency seems implausible.

For these reasons, this study suggested a noncompensatory approach, PB-NM, as an alternative and examined how well the passage-related and passage-independent characteristics of items are interpreted and recovered under each model. The simulation and real data analyses results demonstrate that the PB-NM effectively captures the variability in item-specific dependencies on passage-related and passage-independent proficiencies in a similar way to the bifactor model, albeit through a difficulty-based, as opposed to discrimination-based, representation. Further, in terms of number of model parameters (and effective number of parameters in a Bayesian analysis), the PB-NM model is less complex than the bifactor model and, thus, less prone to overfit, an issue seen especially in cases where there is a small number of items for each passage, resulting in a larger number of passage-specific dimensions in this study. The real data analyses results also demonstrate that the PB-NM confirms the theoretically expected relationships between item types and difference in difficulties across the components, supporting its use in place of the bifactor model when seeking to understand the passage-related effects of items.

We can conceive of several possible applications where the PB-NM may be useful. One application is differential item functioning (DIF) assessment, where the possibility of understanding DIF in a passage-based item may be enhanced by thinking of the two different sources of item difficulty as also reflecting distinguishable sources of DIF. For example, it is possible that DIF in a passage-based item might occur either due to (1) a greater difficulty in one respondent group of processing the passage sufficiently well, or (2) a greater difficulty in understanding the question posed by the item (for reasons unrelated to passage processing). Moreover, as multiple items measure the same passage-based factor, it would also be possible to study differential bundle functioning in a way that separates passage-related from passage-independent influences, potentially increasing statistical power for such studies. A second area of application relates to the cognitive modeling of subprocess difficulty parameters. Explanatory models of item difficulty can naturally be applied to each subprocess within the proposed model, allowing for a better understanding of the contributing factors to item difficulty. Along these lines, models of item generation for fixed passages seemingly may also be more plausible under a noncompensatory MIRT framework.

We can view the noncompensatory approach as also being in line with the emerging interests in the application of sequential tree-based models as measurement models (e.g., De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). Although the PB-NM proposed in this study does not explicitly require sequential processing, its noncompensatory features are consistent with such representations. In this respect, I believe the proposed model may also find useful applications in simulation based measurement tasks where common distinguishable proficiencies are measured within an item across simulation tasks, and where the capability to identify the relative difficulties of the component processes with a particular simulation task is useful. Another application would be to tests for which cognitive diagnostic models (CDMs) are applied, due to the noncompensatory interactions assumed among skills, albeit here with continuous as opposed to discrete proficiencies being invoked. In such contexts, the ability to attach a distinct difficulty to the process associated with each invoked skill would lead to an enhanced understanding about the psychometric functioning of the item.

Future work might not only consider applications such as those described above, but also explore more carefully the conditions necessary to make noncompensatory models more estimable with passage-based tests. It has been clear from previous studies of noncompensatory MIRT models (e.g., Bolt & Lall, 2003) that the additional dimensions (in this case, the passage-based dimensions) must be sufficiently strong for the noncompensatory model to be estimable. The results from this study similarly imply that a dimension switching problem can possibly occur when passage-related dimensions in the data are weak, such as due to small amounts of variance in the passage-related factors, few items per passage, or limited variability in the difficulties of items across dimensions. This suggests the model and estimation algorithm proposed in this study should always be applied with appropriate caution, and a further exploration of necessary conditions for applying the model would be useful.

There is naturally also a potential for other forms of multidimensionality to be present in the data, as might be attributed to the variable item types or disproportionate associations among passages, for example. Such additional sources of dimensionality could interfere with the noncompensatory approach, or alternatively become added components to the model. Further simulation work along the lines of Luecht and Ackerman (2018) could prove useful in this regard. We can also further seek to exploit other empirical relationships between compensatory and noncompensatory models. For example, Bolt (2019) demonstrated how a bifactor model provides a potentially convenient alternative to cognitive diagnostic models (CDMs) that model item responses in relation to a potentially large number of skill attributes, despite the likelihood that the items would only measure the most difficult of the required attributes. The possibility of moving back-and-forth between models seemingly allows users to exploit the model type that makes the most sense in a particular context without worrying too much about substantial differences in model fit.

Such an application would be better supported by more detailed study of how well the PB-NM and associated Bayesian estimation algorithm recovers the examinee proficiency parameters. In this study, I did not evaluate proficiency estimation specifically, nor how the PB-NM and bifactor model adjust the standard errors of proficiency estimates to account for passage-related dependence. As this is often an important feature of testlet models, comparing the models in terms of proficiency estimation would also provide useful insights for understanding the relationship between the models.

Moreover, I did not consider the possibility of having lower asymptotes in the models while including lower asymptotes may better represent responses on multiple-choice items. The studied models, thus, could be evaluated in the presence of estimated non-zero lower asymptotes as is often done when multiple-choice items are involved, such as in the real data analysis of this study. Also, the ability to compare the effects of passage related dimensions across items from different passages is subject to scaling decisions in regard to the passage-related factors. This issue, which applies both to the bifactor and PB-NM models, implies that the way in which passage-related difficulty or discrimination is defined and compared to the general factor can be affected by how the latent variance of the passage-related factors is defined. The effects of these scaling decisions will be more pronounced if there is substantial heterogeneity in passage-related variance across passages. It would be interesting to study how scaling decisions influence examining the effects of passage-related factor in different ways for PB-NM and bifactor model. Finally, the real data application in this study is perhaps simplistic in its assumptions of a single general reading comprehension proficiency. It is conceivable that this ability varies depending on the type of reading passage, and that more attention could be given as to how different passage types (e.g., narrative versus expository) might involve statistically distinguishable reading skills.

3 A Mixture IRTree Model for Extreme Response Style: Accounting for Response Process Uncertainty [2]

3.1 Introduction

Response styles have long been recognized as a threat to the validity of measurement. By definition, response styles refer to systematic tendencies to select response categories in ways that are unrelated to the content of item/test (Paulhus, 1991). One of the most frequently observed response styles is extreme response style (ERS), which refers to the tendency to over-select extreme response categories (de Jong et al., 2008; Greenleaf, 1992). ERS is desirable to measure not only because of its frequent and statistically detectable presence but also due to its potential to contaminate the estimation of the content trait. Ignoring the possibility that extreme responses can be due to ERS may yield an under- or over-estimation of the content trait. Moreover, extreme response styles tend to correlate with other respondent characteristics. Specifically, many researchers have found that both individual- and country-level variables such as sociodemographic, personality and cultural characteristics can correlate with ERS (Austin et al., 2006; C. Chen et al., 1995; de Jong et al., 2008; Greenleaf, 1992; Johnson et al., 2005; Meisenberg and Williams, 2008; Naemi et al., 2009; see Van Vaerenbergh and Thomas, 2012 for more details). Such correlations open the potential for ERS to contribute bias to the estimated relationships between the content trait and other criterion variables.

Item response tree (IRTree) models have recently become a popular methodological ap-

[2] This chapter is an adapted version of Kim and Bolt (2021) published in *Educational and Psychological Measurement*.

proach for the measurement of response styles (Böckenholt, 2012; Böckenholt and Meiser, 2017; Jeon and De Boeck, 2016; Khorramdel and von Davier, 2014; Park and Wu, 2019; Plieninger and Meiser, 2014). Central to the application of such models is the assumption of a sequential process by which respondents arrive at chosen response categories. IRTree models (Böckenholt, 2012; De Boeck and Partchev, 2012; Jeon and De Boeck, 2016) characterize the underlying response processes associated with response selection using a sequential decision tree structure where item response theory (IRT) measurement models represent the outcomes at each decision node. For example, responses to a four-category item with categories "1 = Strongly Disagree, 2 = Disagree, 3 = Agree, and 4 = Strongly Agree" might be specified as a two-stage selection process involving three decision nodes. At the first decision node (i.e., first stage), the respondent chooses whether to agree or disagree with the item based on their content trait, while at a second stage, the respondent decides on the extremity of agreement/disagreement based on their ERS. The outcome at each node is binary, with respondents moving to the second node only if choosing to disagree at the first node, and moving to the third node only if choosing to agree at the first node. By individually aligning the content and response style traits with different response nodes, the IRTree model separates the sources of information contributing to estimates of the content and response style traits, thus, making each easier to estimate.

An implicit assumption of IRTree models is that a single response process is assumed for all respondents. That is, both the decision tree structure and the nature of the underlying traits involved at each decision node are assumed to be the same across respondents. This assumption, however, might be questioned to the extent that different respondents could choose the same response category for different reasons. We might anticipate, for example, that for a certain subset of respondents, selection of the extreme categories at the second stage is not due to a response style, but is instead a further manifestation of their underlying content trait (as the item/test developer presumably intends). In this respect, this chapter proposes a mixture IRTree approach that attends to the possible heterogeneity in response processes across respondents. Through a simulation study, I evaluate the mixture IRTree model and examine how the mixture approach accounts for the presence of a mixture of respondents exhibiting different response processes. The mixture IRTree model is also applied to actual data to demonstrate the presence of response process heterogeneity in self-report rating scale items, and further illustrate how the mixture IRTree model can accommodate the possibility that different traits may be relevant for different respondents in explaining their response category selection.

3.2 Model

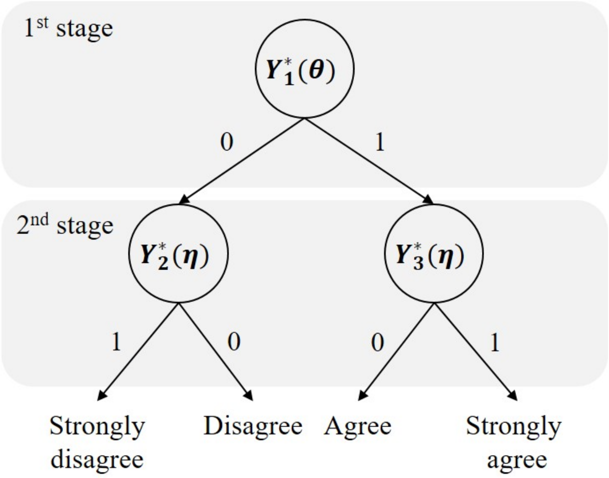
In this section, a mixture IRTree model attending to a mixture of respondents resorting to different response processes in responding to 4-point Likert scale items is presented. To develop a mixture model, two different IRTree models are considered: ERS and ordinal (ORD) IRTree models. Below I describe statistical representations of the two IRTree models and how these two models are combined into a latent mixture model.

3.2.1 IRTree Model for Extreme Response Style: The ERS IRTree Model

As briefly described in the introduction, an IRTree model that accounts for ERS for responses to a 4-point Likert scale item (i.e., ERS IRTree model) is commonly specified as a two-stage process with three decision nodes involved. Figure 3.1 shows a graphical representation of the decision tree structure for the model. A common approach to translating an IRTree model to a statistical representation makes use of *pseudo-items* that correspond to the binary outcome at each decision node. I denote the outcome at decision node k for respondent j on item i as pseudo-item Y^*_{ijk}, where a value of 1 typically corresponds to the decision reflecting a higher level of the trait underlying that node. Consequently, for the IRTree model in Figure 3.1, an "Agree" decision at the first node is coded as $Y^*_{ij1} = 1$ while a "Disagree" decision is coded as $Y^*_{ij1} = 0$ as a respondent possessing a high level of content trait (which underlies the first node) would tend to "agree" to the statement. At Nodes 2 and 3, $Y^*_{ij2} = 1$ and $Y^*_{ij3} = 1$ correspond to the selection of extreme responses (i.e., "Strongly agree" or "Strongly disagree"), and $Y^*_{ij2} = 0$ and $Y^*_{ij3} = 0$ to selection of non-extreme responses (i.e., "Agree" or "Disagree") because ERS underlies the second and third nodes. This implies that responses to each category in Figure 3.1 can be recoded into three pseudo-items that correspond to the three decision nodes. For example, a response of "2 = Disagree" can be recoded to 0 for both Y^*_{ij1} and Y^*_{ij2}, and missing ("NA") for Y^*_{ij3} (as respondents who disagree at Node 1 will not go through Node 3).

Figure 3.1

Illustration of an Extreme Response Style (ERS) Item Response Tree (IRTree) Model for Responses to a 4-Point Likert-Type Scale Item



Given this coding of pseudo-items, the probability of an outcome at a decision node is represented using an IRT model. In this study, I assume a two-parameter logistic (2PL) model for all nodes. For decision node $k = 1$, the probability of an outcome would be

$$P(Y_{ij1}^* = 1) = \frac{\exp(a_{i1}\theta_j + b_{i1})}{1 + \exp(a_{i1}\theta_j + b_{i1})} \tag{3.1}$$

where θ_j denotes the content trait, while for nodes $k = 2$ and 3, it would be

$$P(Y_{ijk}^* = 1) = \frac{\exp(a_{ik}\eta_j + b_{ik})}{1 + \exp(a_{ik}\eta_j + b_{ik})} \tag{3.2}$$

where η_j denotes the extreme response style trait. Note that the models are expressed using the slope-intercept parameterization where the a_{ik} denotes item discrimination parameters and b_{ik} denotes intercept (difficulty-related) parameters. The a_{ik} and b_{ik} are allowed to differ across nodes, with a_{ik} assumed to be positive. An assumption of local independence across nodes makes the probability of selecting a particular category m equal to the product of probabilities of responses to pseudo-items across the three nodes:

$$\begin{aligned}P(Y_{ij} = m) =& P(Y^*_{ij1} = y^*_{ij1})P(Y^*_{ij2} = y^*_{ij2})P(Y^*_{ij3} = y^*_{ij3})\\ =& \frac{\exp[y^*_{ij1}(a_{i1}\theta_j + b_{i1})]}{1 + \exp(a_{i1}\theta_j + b_{i1})} \times \left[\frac{\exp[y^*_{ij2}(a_{i2}\eta_j + b_{i2})]}{1 + \exp(a_{i2}\eta_j + b_{i2})}\right]^{1-y^*_{ij1}}\\ &\times \left[\frac{\exp[y^*_{ij3}(a_{i3}\eta_j + b_{i3})]}{1 + \exp(a_{i3}\eta_j + b_{i3})}\right]^{y^*_{ij1}}\end{aligned} \tag{3.3}$$

where y^*_{ijk} for each node k combine to render a final response of m. The probability of selecting “2=Disagree” category, for instance, would be

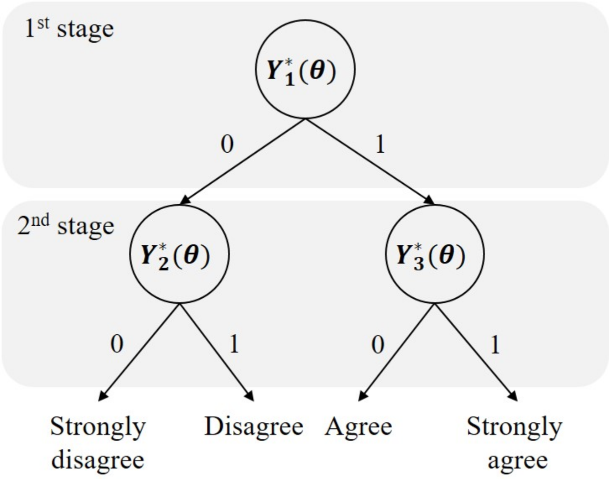
$$P(Y_{ij} = 2) = P(Y^*_{ij1} = 0)P(Y^*_{ij2} = 0) = \frac{1}{1 + \exp(a_{i1}\theta_j + b_{i1})} \times \frac{1}{1 + \exp(a_{i2}\eta_j + b_{i2})} \tag{3.4}$$

3.2.2 An Alternative Model: The Ordinal (ORD) IRTree Model

As an alternative to the ERS IRTree model, I consider a model that assumes a similar sequential process, but where the content trait underlies decisions made at all three nodes. As a result, the general tree presentation in Figure 3.1 still applies, but the decisions made at the second stage (Nodes 2 and 3) are assumed to be influenced by θ (content trait) as opposed to η (ERS). Figure 3.2 illustrates the resulting Ordinal (ORD) IRTree model now having a binary outcome at Decision Nodes 2 and 3 that implies a selection of the higher of the two successive categories (as opposed to the most extreme of the options). Therefore, the binary outcome at the second node is reversed compared with the ERS model in Figure 3.1. Specifically, “Disagree” corresponds to $Y^*_{ij2} = 1$ and “Strongly disagree” corresponds to $Y^*_{ij2} = 0$ in this ORD model (as a respondent with a higher level of the content trait is more likely to choose “Disagree” than “Strongly Disagree”).

Figure 3.2

Illustration of an Ordinal (ORD) Item Response Tree (IRTree) Model for Responses to a 4-Point Likert-Type Scale Item



The outcome at each decision node is consequently modeled as

$$P(Y^*_{ijk} = 1) = \frac{\exp(a_{ik}\theta_j + b_{ik})}{1 + \exp(a_{ik}\theta_j + b_{ik})} \tag{3.5}$$

for $k = 1, 2, 3$ where θ_j denotes the content trait for respondent j and a_{ik}s are constrained to be positive. The item parameters a_{ik} and b_{ik} are separately estimated for each node as in the ERS model.

Note that the same pseudo-items created for the ERS model can also be applied in fitting the ORD model. However, because the outcome for Node 2 is reversed under the ORD model compared with the ERS model (i.e., $Y^*_{ij2} = 1$ under the ERS model corresponds to the lower category “Strongly Disagree” rather than the higher score category “Disagree”), the outcome for the second node can be specified as

$$P(Y^*_{ij2} = 1) = 1 - \frac{\exp(a_{i2}\theta_j + b_{i2})}{1 + \exp(a_{i2}\theta_j + b_{i2})} = \frac{1}{1 + \exp(a_{i2}\theta_j + b_{i2})} = \frac{\exp(-a_{i2}\theta_j - b_{i2})}{1 + \exp(-a_{i2}\theta_j - b_{i2})} \tag{3.6}$$

for the ORD model using the pseudo-item created under the ERS model. In this way, each of the ERS and ORD IRTree models can be fitted to the same pseudo-items. Similar to the ERS IRTree model, an assumption of local independence across nodes makes the probability of selecting category m under the ORD model equivalent to

$$\begin{aligned} P(Y_{ij} = m) =& \frac{\exp[y^*_{ij1}(a_{i1}\theta_j + b_{i1})]}{1 + \exp(a_{i1}\theta_j + b_{i1})} \times \left[\frac{\exp[y^*_{ij2}(-a_{i2}\theta_j - b_{i2})]}{1 + \exp(-a_{i2}\theta_j - b_{i2})}\right]^{1-y^*_{ij1}} \\ & \times \left[\frac{\exp[y^*_{ij3}(a_{i3}\theta_j + b_{i3})]}{1 + \exp(a_{i3}\theta_j + b_{i3})}\right]^{y^*_{ij1}} \end{aligned} \tag{3.7}$$

where y^*_{ij1} refers to the pseudo-item responses as defined under the ERS tree in Figure 3.1.

3.2.3 A Mixture of ERS and ORD IRTree Models

The ORD model is naturally a competitor to the ERS model and could be statistically compared with the ERS IRTree model. However, when both models are viewed as applicable across a population of respondents, a mixture IRTree model can be formulated in which each of the ERS and ORD IRTree models defines a latent class in the mixture. Under a mixture representation, each respondent is assumed to have a latent membership in either the ERS or

ORD class across all item responses. As can be seen in Figures 3.1 and 3.2, the two classes are distinguished by whether the decisions at Stage 2 (Nodes 2 and 3) are affected by the content trait θ or an extreme response style trait η. In order to link metrics of the content trait across classes, I constrain the ORD and ERS IRTree models to share a common latent trait θ (i.e., content trait) at the first node and have identical item parameters for that first pseudo-item across classes. The item parameters for Nodes 2 and 3 are allowed to vary across classes and are separately estimated for the two classes. Consequently, the mixture IRTree model can be written as

$$
\begin{aligned}
P(Y_{ij} = m) = {} & \frac{\exp[y^*_{ij1}(a_{i1}\theta_j + b_{i1})]}{1 + \exp(a_{i1}\theta_j + b_{i1})} \\
& \times \left\{ \left[\frac{\exp[y^*_{ij2}(a_{i2,ERS}\eta_j + b_{i2,ERS})]}{1 + \exp(a_{i2,ERS}\eta_j + b_{i2,ERS})} \right]^{1-y^*_{ij1}} \left[\frac{\exp[y^*_{ij3}(a_{i3,ERS}\eta_j + b_{i3,ERS})]}{1 + \exp(a_{i3,ERS}\eta_j + b_{i3,ERS})} \right]^{y^*_{ij1}} \right\}^{2-z_j} \\
& \times \left\{ \left[\frac{\exp[y^*_{ij2}(-a_{i2,ORD}\theta_j - b_{i2,ORD})]}{1 + \exp(-a_{i2,ORD}\theta_j - b_{i2,ORD})} \right]^{1-y^*_{ij1}} \left[\frac{\exp[y^*_{ij3}(a_{i3,ORD}\theta_j + b_{i3,ORD})]}{1 + \exp(a_{i3,ORD}\theta_j + b_{i3,ORD})} \right]^{y^*_{ij1}} \right\}^{z_j-1}
\end{aligned}
\tag{3.8}
$$

where z_j denotes the class membership parameter of respondent j (1 = ERS class, 2 = ORD class), y^*_{ij1} denotes pseudo-item responses recoded under the ERS model, $a_{ik,ERS}$ and $b_{ik,ERS}$ for $k = 2, 3$ represent item parameters of Nodes 2 and 3 for the ERS class, and $a_{ik,ORD}$ and $b_{ik,ORD}$ for the ORD class. We can observe from the right-hand side of the Equation 3.8 that the first part (i.e., the probability for the first node) stays the same for both classes but either the second or the third part drops out depending on the class membership z_j a respondent has. For instance, when a respondent belongs to the ORD class ($z_j = 2$), the second part drops out and the third part stays in the equation as the exponents $2 - z_j$ and $z_j - 1$, respectively, become 0 and 1. In contrast, when a respondent is in the ERS class ($z_j = 1$), the third part drops out and the second part stays in the equation.

Table 3.1 summarizes the pseudo-item outcomes that correspond to each item category response and the probability of responses at each node for each model (class) in the mixture model. I present the pseudo-item outcomes created under the ERS model and, therefore, use Equation 3.6 for the probability at Node 2 for the ORD model.

Table 3.1
A Summary of Information for the Mixture Item Response Tree (IRTree) Model

Node	Psuedo-item for category m (y^*_{ijk})				Class	Model
(trait)	$m = 1$	$m = 2$	$m = 3$	$m = 4$	membership (z_j)	$[P(Y^*_{ijk} = y^*_{ijk})]$
$Y^*_1(\theta)$	0	0	1	1	Both(1,2)	$\frac{\exp[y^*_{ij1}(a_{i1}\theta_j + b_{i1})]}{1+\exp(a_{i1}\theta_j + b_{i1})}$
$Y^*_2(\eta)$	1	0	NA	NA	ERS(1)	$\frac{\exp[y^*_{ij2}(a_{i2,ERS}\eta_j + b_{i2,ERS})]}{1+\exp(a_{i2,ERS}\eta_j + b_{i2,ERS})}$
$Y^*_2(\theta)$	1	0	NA	NA	ORD(2)	$\frac{\exp[y^*_{ij2}(-a_{i2,ORD}\theta_j - b_{i2,ORD})]}{1+\exp(-a_{i2,ORD}\theta_j - b_{i2,ORD})}$
$Y^*_3(\eta)$	NA	NA	0	1	ERS(1)	$\frac{\exp[y^*_{ij3}(a_{i3,ERS}\eta_j + b_{i3,ERS})]}{1+\exp(a_{i3,ERS}\eta_j + b_{i3,ERS})}$
$Y^*_3(\theta)$	NA	NA	0	1	ORD(2)	$\frac{\exp[y^*_{ij3}(a_{i3,ORD}\theta_j + b_{i3,ORD})]}{1+\exp(a_{i3,ORD}\theta_j + b_{i3,ORD})}$

3.3 Simulation Analyses

The mixture IRTree model and its estimation using a fully Bayesian estimation algorithm are evaluated with simulated data. I focus on evaluating how well the model identifies the mixture of classes (both at respondent and sample levels) and recovers item and respondent parameters. Below I describe data simulation process/conditions and how the generated data are analyzed.

3.3.1 Data Simulation

The ERS and ORD IRTree models in Equations 3.3 and 3.7, respectively, are used to generate response patterns for respondents in Classes 1 (ERS class) and 2 (ORD class). Responses for a total of 1,000 respondents to 15 four-response category items are generated. The proportion of respondents in each class is systematically varied: (P_1, P_2) = (1.0, 0.0), (0.7, 0.3), (0.5, 0.5), (0.7, 0.3), and (0.0, 1.0) as a simulation factor, where P_1 and P_2, respectively, denote the proportion of respondents in the ERS and ORD classes. Ten data sets are generated for each condition for replication purposes; accordingly, 50 data sets (10 replications $\times$ 5 conditions) are generated in total. For each respondent, item category responses are simulated as outcomes of a sequential response process corresponding to the IRTree model of their respective class. Then the category responses are recoded into pseudo-items as defined under the ERS IRTree model. The overall data generation process can be summarized in the following steps:

Step 1. Generate person parameter θ_j (content trait) and η_j (ERS) for 1,000 respondents independently from $Normal(0, 1)$, assuming that θ_j and η_j are uncorrelated.

Step 2. Generate item parameters a_{ikc} and b_{ikc} across three nodes ($k = 1, 2, 3$) for 15 items in each of the two classes ($c = 1(ERS), 2(ORD)$), respectively, as $Uniform(0.5, 2)$ and $Uniform(-3, 3)$. The item parameters for the first node were generated to be identical across the two classes ($a_{i1,ERS} = a_{i1,ORD}$, $b_{i1,ERS} = b_{i1,ORD}$) while the parameters for the second and third nodes were independently generated for each class.

Step 3. Assign class membership parameters (z_j) for respondents according to the mixture proportion condition (i.e., $[P_1, P_2]$) being considered. I assigned the first $100P_1$% of the respondents to Class 1 (ERS class) and the rest to Class 2 (ORD class).

Step 4. Calculate the probability of each respondent selecting category $m (= 1, 2, 3, 4)$ using Equation 3.8. For instance, for a respondent in ERS class ($z_j = 1$), the probability of selecting category m is calculated by plugging in the generated parameter values for $a_{ij,ERS}$ and $b_{ij,ERS}$ and pseudo-item outcome values corresponding to category m for y^*_{ijk} in this equation: $\frac{\exp[y^*_{ij1}(a_{i1}\theta_j + b_{i1})]}{1+\exp(a_{i1}\theta_j + b_{i1})} \left[\frac{\exp[y^*_{ij2}(a_{i2,ERS}\eta_j + b_{i2,ERS})]}{1+\exp(a_{i2,ERS}\eta_j + b_{i2,ERS})}\right]^{1-y^*_{ij1}} \left[\frac{\exp[y^*_{ij3}(a_{i3,ERS}\eta_j + b_{i3,ERS})]}{1+\exp(a_{i3,ERS}\eta_j + b_{i3,ERS})}\right]^{y^*_{ij1}}$. Four probability values (corresponding to four response categories) for each respondent are calculated.

Step 5. Generate multinomial responses from 1,000 respondents to 15 items, using the probabilities calculated in the previous step.

Step 6. Transform the categorical responses to pseudo-items based on the ERS tree structure, as shown in Table 3.1. (Recall that the ORD tree structure can be fitted to ERS tree pseudo-items by forcing the discrimination parameter at the second node to be negative as opposed to positive; see Equation 3.6.) The final data set consequently has binary responses from 1,000 respondents to 45 pseudo-items (15 items × 3 nodes) with the irrelevant pseudo-items coded as missing.

Step 7. Repeat Steps 5 and 6 within each condition to generate 10 data sets for each mixing proportion condition.

Step 8. Repeat Steps 3 through 7 for each of the mixing proportion conditions.

3.3.2 Analyses

I fit the IRTree mixture model involving ERS and ORD classes to each of the generated data sets using a Bayesian (Markov chain Monte Carlo) estimation algorithm using JAGS (Just Another Gibbs Sampler) 4.3.0 (Plummer, 2017). To run JAGS from the R software (R Core Team, 2019), the *jagsUI* package (Kellner, 2019) is used. The prior distributions for the item parameters a_{ikc} and b_{ikc} are, respectively, $logNormal(0, 1)$ and $Normal(0, 1)$. For person parameters, θ_j and η_j are each assumed to independently follow a prior distribution of $Normal(0, 1)$ while the class membership parameter z_j is assumed to be $Categorical(p_1, p_2)$ with $Dirichlet(\alpha_1, \alpha_2)$ as a prior for the vector of hyperparameters (p_1, p_2). I set $\alpha_1 = \alpha_2 = 1$ to make the prior uniformly distributed and noninformative. For each simulated data set, 10 chains are run where for each chain the total number of iterations is set to 25,100. Standard convergence criteria (i.e., Gelman–Rubin R^2; Gelman and Rubin, 1992) are used to support the successful convergence. The first 100 iterations are used for adaptation and a subsequent 5,000 iterations are discarded as burn-in. Every 10th subsequent value in the simulated chains are retained (i.e., thinning interval = 10 iterations), implying that a total of 2,000 iterations are used from each of the 10 chains to produce posterior distributions of the model parameters. The mean values of the univariate posterior distributions are used as parameter estimates.

To compare the mixture IRTree model against the use of a single IRTree model, I also separately fit the ERS and ORD IRTree models to the same datasets to examine whether the mixture model emerges as superior in the presence of two classes. The same prior distributions as used in the mixture model are applied for the corresponding parameters under each single

IRTree model. Due to the reduced complexity of these models, five chains are run for each analysis and a total of 15,100 iterations for each chain. The first 100 and 5,000 iterations are discarded as adaptive and burn-in iterations, respectively. The resulting posterior distributions are constructed from the 10,000 post burn-in iterations again using a thinning interval of 10, implying a total of 1,000 iterations from five respective chains for determination of parameter estimates.

3.4 Simulation Results

3.4.1 Model Fit Comparisons

I first compared the fit of the mixture IRTree model with that of the single ERS and ORD IRTree models. The deviance information criteria (DIC) for the models, obtained from the first simulated data set, are reported in Table 3.2. The DIC is calculated as the sum of the mean deviance to a penalty based on the complexity of the model (the effective number of parameters, denoted as pD). As expected, for the data sets that only contained respondents from one of the two classes, that is $(P_1, P_2) = (1.0, 0.0), (0.0, 1.0)$, the IRTree model that corresponds to the correct class returned the lowest DIC value indicating the best model fit, while the mixture model returned the lowest DIC value for all conditions involving a mixture of the two classes. The same pattern of findings was observed across all 10 of the data sets, suggesting that the mixture IRTree model correctly emerges as superior in the presence of conditions in which both the ERS and ORD classes are present for different respondent subpopulations.

Table 3.2

Deviance Information Criterion (DIC) Results for the Mixture, ERS, and ORD IRTree Models for Five Different Mixture Proportion Conditions, First Simulated Data Set for Each Condition

Fitted model	Class mixing proportion condition (P_1, P_2)									
	(1.0, 0.0)		(0.7, 0.3)		(0.5, 0.5)		(0.3, 0.7)		(0.0, 1.0)	
	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC
Mixture	2529	29107	2583	**28948**	2402	**29051**	2169	**28433**	1359	27519
ERS	2308	**28898**	2439	30932	2486	31718	2406	31391	2719	30228
ORD	1249	31726	1256	32422	1264	32125	1195	30662	1201	**27374**

Note. The smallest DIC values for each condition is in boldface. ERS = Extreme response style; ORD = Ordinal; pD = Effective number of parameters; AIC = Akaike Information Criterion.

3.4.2 Estimation of Latent Proportions and Classification Accuracy

The ability of the mixture IRTree model to correctly capture the mixture of respondents in the data can also be inspected by looking at how well the model estimates the mixing proportion parameters (p_1, p_2) and the respondent class memberships (z_j). I evaluated recovery in terms of the bias and root mean square errors (RMSEs) of the estimates of the latent proportions. Specifically,

$$Bias = E(\hat{p}_{1r}) - p_1 = \frac{1}{rep}\sum_{r=1}^{rep}\hat{p}_{1r} - p_1 \tag{3.9}$$

$$RMSE = \sqrt{E[(\hat{p}_{1r}) - p_1)^2]} = \sqrt{\frac{1}{rep}\sum_{r=1}^{rep}(\hat{p}_{1r} - p_1)^2} \tag{3.10}$$

where $\hat{p}_{1r}$ denotes the estimate of the proportion parameter p_1 obtained from the rth simulation data set, and rep denotes the total number of replications for each condition (in this case 10).

As can be seen in Table 3.3, the bias and RMSE are all very small (less than 0.01) using the fully Bayesian estimation approach, indicating that the mixture model estimates the true proportion of respondents in each class accurately. At the respondent level, the estimate of class membership ($\hat{z}_j$) lies between 1 and 2, and reflects the probability of belonging to each class. For example, if a respondent's estimated class membership is 1.3, this means that the probability of belonging to Class 2 (ORD class) for the individual is 0.3 and to Class 1 (ERS class) is 0.7. Respondents are thus assigned to Class 1 if their $\hat{z}_j$ is smaller than or equal to 1.5 and Class 2 if their $\hat{z}_j$ is larger than 1.5. I evaluate the accuracy of such a classification by calculating the proportion of correctly assigned respondents to each class (i.e., hit rate). The average hit rate across the 10 replicated data sets within each condition is presented in the right two columns of Table 3.3. The hit rates are larger than 90% for all conditions, although accuracy decreases slightly as the proportion of respondents in the class decreases. For instance, the hit rate for the ERS class is 99.97% when all the respondents in the data are in the ERS class, but decreases to 91.37% when the proportion of respondents in the ERS class reduces to 0.3. These overall results suggest that the proposed modeling approach assigns the respondents to their correct classes with high accuracy provided the number of respondents in the class is sufficiently large.

3.4.3 Recovery of Pseudo-Item Parameters

I also examined how well the mixture model recovers the pseudo-item parameters. Table 3.4 displays the bias and RMSE of the item parameters a_{ikc} and b_{ikc} for each class ($c = 1, 2$) averaged over items and nodes. The values derived from the single IRTree models are also

Table 3.3

Bias and Root Mean Square Error (RMSE) of the Estimated Latent Proportion $\hat{p}_1$ and Classification Accuracies (Average Hit Rates) Across Five Different Mixing Proportion Conditions

Condition (P_1, P_2)	Posterior latent proportion $\hat{p}_1$		Average hit rate (%)	
	Bias	RMSE	P($\hat{z}_j \leq$1.5$|z_j = 1$)	P($\hat{z}_j$>1.5$\|z_j = 2$)
(1.0, 0.0)	-0.003	0.003	99.97	-
(0.7, 0.3)	-0.006	0.009	96.51	92.53
(0.5, 0.5)	-0.002	0.009	94.64	96.14
(0.3, 0.7)	-0.003	0.008	91.37	97.89
(0.0, 1.0)	0.002	0.002	-	100.00

presented as criteria (baseline) for evaluation. The mixture model produces nearly the same levels of bias and RMSE for pseudo-item discrimination and difficulty parameters as those of the single IRTree models when the proportion of respondents under the corresponding model is 1. For the intermediate mixing proportion conditions, the mixture model always produces smaller bias and RMSE compared with the single IRTree models, suggesting that the mixture model provides a superior estimation of the pseudo-item parameters. In addition, the mixture model seems to recover well the item parameters for a given class even when there is only a small proportion of respondents in the class. For instance, when the proportion of respondents in ERS class is only 0.3, the bias and RMSE of $a_{ik1}(a_{ERS})$ estimates produced from the mixture model are, respectively, 0.133 and 0.263, whereas the ERS model produced larger values, presumably because, under the ERS model, information from respondents who are in the ORD class are also applied in estimating the item parameters.

Table 3.4
Bias and Root Mean Square Error (RMSE) of the Item Parameter Estimates Across Mixture, ERS and ORD IRTree Models by Mixing Proportion Condition.

Parameter Estimate	Condition (P_1, P_2)	Bias[a]			RMSE		
		MIX	ERS	ORD	MIX	ERS	ORD
$\hat{a}_{ik1}(\hat{a}_{ERS})$	(1.0, 0.0)	0.072	0.071	-	0.184	0.183	-
	(0.7, 0.3)	0.082	0.253	-	0.208	0.328	-
	(0.5, 0.5)	0.097	0.445	-	0.234	0.504	-
	(0.3, 0.7)	0.135	0.547	-	0.324	0.605	-
	(0.0, 1.0)	0.323	0.483	-	0.365	0.551	-
$\hat{a}_{ik2}(\hat{a}_{ORD})$	(1.0, 0.0)	0.339	-	0.660	0.398	-	0.704
	(0.7, 0.3)	0.107	-	0.542	0.296	-	0.595
	(0.5, 0.5)	0.093	-	0.418	0.258	-	0.467
	(0.3, 0.7)	0.086	-	0.283	0.222	-	0.343
	(0.0, 1.0)	0.064	-	0.065	0.185	-	0.184
$\hat{b}_{ik1}(\hat{b}_{ERS})$	(1.0, 0.0)	0.071	0.072	-	0.155	0.156	-
	(0.7, 0.3)	0.076	0.370	-	0.168	0.417	-
	(0.5, 0.5)	0.092	0.558	-	0.213	0.603	-
	(0.3, 0.7)	0.133	0.720	-	0.263	0.762	-
	(0.0, 1.0)	0.927	0.998	-	0.949	1.035	-
$\hat{b}_{ik2}(\hat{b}_{ORD})$	(1.0, 0.0)	0.971	-	1.007	0.998	-	1.037
	(0.7, 0.3)	0.152	-	0.752	0.274	-	0.784
	(0.5, 0.5)	0.128	-	0.610	0.257	-	0.646
	(0.3, 0.7)	0.102	-	0.459	0.198	-	0.496
	(0.0, 1.0)	0.079	-	0.080	0.158	-	0.159

Note. MIX = Mixture IRTree model; ERS = Extreme response style IRTree model; ORD = Ordinal IRTree model.

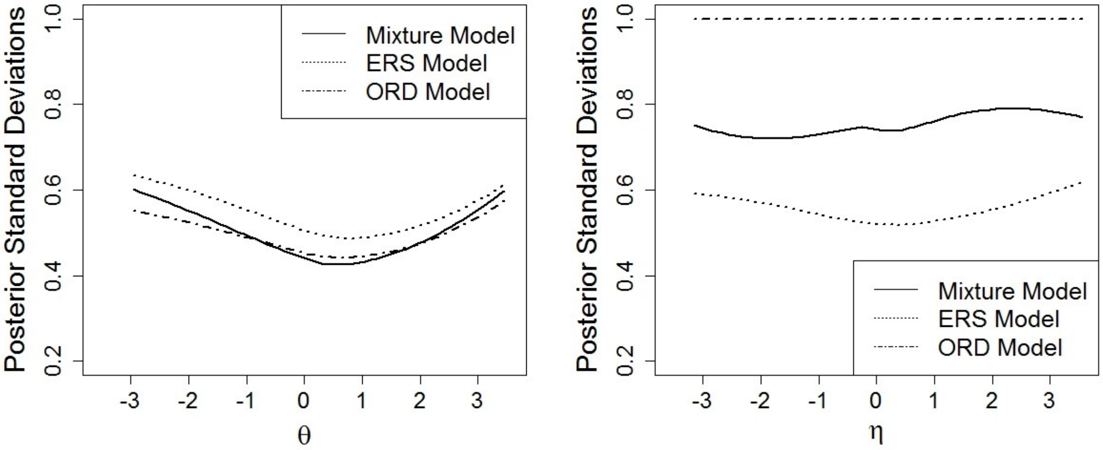
[a] The *absolute values* of bias are averaged over items and nodes.

3.4.4 Precision (Posterior Standard Deviations) of Respondent Parameter Estimates

Appendix C displays results in terms of the bias and RMSE of respondent parameter recovery (for both θ and η). Importantly, the use of the mixture IRTree model is seen to result in a recovery of respondent θ estimates that is equivalent to that observed when applying the correct model to the respondent's data. In other words, under the mixture IRTree model, the recovery of θ for ORD respondents is as good as it is when applying the single ORD IRTree model, and likewise, the recovery for ERS respondents is as good as when applying the single ERS IRTree model. Similarly, for the η estimates, recovery for ERS respondents under the mixture IRTree model is as good as when applying the ERS model. (Note that recovery of the η for ORD respondents is consistently poor under both models, as the items do not measure η for this class of respondents.) As a result, the application of a mixture approach seemingly renders accurate trait estimates with respect to the relevant traits for the respondent's actual response process.

Of particular interest in the current application, however, is the way in which the mixture approach represents the precision of the respondent parameter estimates. As suggested in the introduction, an anticipated consequence of applying the mixture IRTree model is that it will appropriately account for the uncertainty present in respondent parameter estimates due to the uncertainty of the respondent's response process. As seen in Figure C2 in Appendix C, both the mixture and ERS models produce equivalently large errors for many of the η estimates. One of the differences between the mixture and ERS models is that the mixture model attends to the uncertainty of the estimates due to the uncertain response process whereas the ERS model does not. Figure 3.3 displays kernel-smoothed functions of the relationship between the respondent parameters and their corresponding posterior standard deviations (the Bayesian equivalent of standard errors of estimates). The left plot shows the posterior standard deviations for θ, the right plot for η. For the content trait θ, it can be observed from the left panel in Figure 3.3 that the mixture model produces lower posterior standard deviations than those from the ERS model across all levels of θ, a result that can be attributed to the capacity of the mixture model to extract additional information about θ from Nodes 2 and 3 when the respondent is in the ORD class. At the same time, the mixture model generally produces higher posterior standard deviations than those of the ORD model, a result that is due to some respondents actually being in the ERS class, as well as the uncertainty regarding class membership for many of the respondents.

Figure 3.3
Average Posterior Standard Deviations (PSDs) of θ and η in Relation to the Parameters, for Mixture and Single IRTree Models, for (P_1, P_2) = (0.5, 0.5) Condition



For the posterior standard deviations of η, I initially note that because only the mixture and ERS models actually estimate η, the result for the ORD model is a constant value of 1. This reflects that from a Bayesian perspective, the posterior standard deviation of η would be 1, the standard deviation of the prior, for all respondents in the ORD class as the data provide no information about η. The right panel of Figure 3.3 further shows that the mixture model returns higher posterior standard deviations across all levels of η in comparison to the ERS model. This result is as expected for two primary reasons: First, the mixture model produces large posterior standard deviations of η for those respondents who belong to the ORD class (where η is unmeasured), and second, the uncertainty of class membership makes the estimates of η less precise even for those belonging to the ERS class.

To evaluate the accuracy of the precision of estimates, I examined the proportion of the 95% credible intervals that contain the true parameter values for each parameter under each modeling approach, as shown in Table 3.5. From the leftmost columns of the table, it is shown that the ERS and mixture models provide accurate 95% credible intervals for the θ parameters, indicating that the mixture model accurately reflects the estimation of θ; however, the intervals under the ORD model are consistently in the lower 90s, suggesting inaccuracy. Recalling from Figure 3.3 that the posterior standard deviations were consistently lower for the mixture model than the ERS model, it would thus appear that the credible intervals for the mixture model are best (probably narrower intervals than for the ERS model with similar accuracy).

Table 3.5

The Proportion of 95% Credible Intervals for θ and η That Contain True Parameters, for the Mixture, ERS, and ORD IRTree Models Across Mixing Proportion Conditions

Condition	θ			η		
(P_1, P_2)	MIX	ERS	ORD	MIX	ERS	ORD[a]
(1.0, 0.0)	0.952	0.952	0.938	0.940	0.940	0.944
(0.7, 0.3)	0.951	0.950	0.913	0.942	0.835	0.944
(0.5, 0.5)	0.954	0.953	0.919	0.945	0.718	0.944
(0.3, 0.7)	0.954	0.952	0.915	0.941	0.688	0.944
(0.0, 1.0)	0.953	0.953	0.952	0.944	0.830	0.944

Note. MIX = Mixture IRTree model; ERS = Extreme response style IRTree model; ORD = Ordinal IRTree model.

[a] 0.025th and 0.975th quantiles of normal distribution are used for the interval.

From the rightmost three columns in Table 3.5, it is likewise seen that the intervals derived from the mixture model mostly include η (about 94%) whereas the ERS model more frequently returns rates much lower, especially in the presence of a near equal split in class sizes. Thus, by attending to the uncertainty of the η estimates, the mixture model appears to produce more accurate credible intervals than the ERS model. In summary, when a mixture is simulated to be present, it is only by modeling the data as a mixture that we can obtain accurate estimates and credible intervals for the respondent parameters.

3.5 Real Data Application

The mixture model was also fitted to actual data to demonstrate the presence of a mixture of respondents in self-report rating scale items as well as to show a real-world illustration of the results seen in the simulation. The data were collected through a survey from the Trends in International Mathematics and Science Study (TIMSS) 2015 (for Grade 8). I used responses from 1,000 randomly sampled respondents from the U.S. administration of the nine items for the Students Like Learning Mathematics (SLM) scale. Each item was rated using four response categories: *disagree a lot, disagree a little, agree a little,* and *agree a lot*. I converted the item scores to pseudo-item responses based on the ERS IRTree model and fitted the mixture model to the data using JAGS (Plummer, 2017) with the same specifications as for the simulation analyses reported above. As in the simulation analyses, the single ERS and ORD IRTree models were also fitted.

I validated the presence of a mixture by comparing the model fit and examining the estimated latent proportions for each class as well as estimated class memberships of individuals in the mixture model. The comparative model fits are provided in Table 3.6. As can be seen, the mixture IRTree model produces the smallest DIC value, suggesting a better fit in comparison to the other two single IRTree models. Also, it is observed from the leftmost two columns in Table 3.7 that the latent proportions for each class estimated from the mixture model are about 0.3 (ERS class) and 0.7 (ORD class), respectively, roughly agreeing with the proportion of respondents actually classified to each class (as can be seen in the rightmost two columns in Table 3.7).

Table 3.6
Deviance Information Criterion (DIC) Results of the Mixture, ERS, and ORD IRTree Models for TIMSS SLM Scale Data

Model	Mixture	ERS	ORD
pD	2636.5	2206.7	1330.3
DIC	**15357.7**	15441.17	15391.48

Note. The *smallest* DIC value is in bolface. TIMSS = Trends in International Mathematics and Science Study; SLM = Students Like Learning Mathematics; ERS = Extreme response style IRTree model; ORD = Ordinal IRTree model; pD = Effective number of parameters.

Table 3.7
Estimated Latent Proportions and the Number of Respondents Classified to ERS and ORD Classes for TIMSS SLM Scale Data

Posterior latent proportions		Number of classified respondents	
ERS class ($\hat{p}_1$)	ORD class ($\hat{p}_2$)	ERS class ($\hat{z}_j \leq$1.5)	ORD class ($\hat{z}_j$>1.5)
0.320	0.680	227	773

Note. TIMSS = Trends in International Mathematics and Science Study; SLM = Students Like Learning Mathematics; ERS = Extreme response style IRTree model; ORD = Ordinal IRTree model.

To illustrate the implications of applying single or mixture IRTree models to the data with a mixture of respondents, some example response patterns and their corresponding trait estimates are provided in Table 3.8. Note that the ERS and mixture IRTree models provide estimates of both θ and η, while the ORD model only provides an estimate of θ (η estimates could also naturally be reported as the mean, 0, standard deviation, 1, as defined by the prior distribution). The mixture model also provides the estimated probability of class membership for each respondent. The estimated probabilities of being in ORD class for Respondents 4 and 141, who selected all 1s or 4s, indicate that it is highly likely that they chose extreme responses due to their low or high content trait θ. The probabilities, however, are not equal to 1 because there still is a possibility that they selected all 1s or 4s due to their ERS. On the other hand, Respondents 244 and 74 have a low probability of belonging to the ORD class, implying that it is highly likely that these responses are influenced by their ERS. Specifically, selecting options only from both ends (1s and 4s) may reflect a high ERS while selecting a middle category from only one direction (all 3s) reflects a low ERS (i.e., tendency to avoid extreme responses). The probabilities for both cases also do not equal to 0, reflecting the possibility of selecting extreme responses due to their content traits.

Table 3.8

Example Response Patterns for TIMSS SLM Scale Data and Corresponding Content and Response Style Trait Posterior Means and Standard Deviations, Derived From the Mixture, ERS, and ORD IRTree Models

ID	Responses	Mixture			ERS		ORD
		Prob(Class2)	$\hat{\theta}$ (PSD(θ))	$\hat{\eta}$ (PSD(η))	$\hat{\theta}$ (PSD(θ))	$\hat{\eta}$ (PSD(η))	$\hat{\theta}$ (PSD(θ))
603	221112211	0.713	-0.956 (0.439)	0.096 (0.863)	-1.411 (0.553)	0.159 (0.250)	-0.866 (0.192)
4	111111111	0.866	-1.789 (0.538)	0.196 (1.087)	-1.423 (0.551)	1.477 (0.556)	-1.922 (0.496)
141	444444444	0.871	1.815 (0.555)	0.212 (1.099)	1.297 (0.581)	1.542 (0.557)	1.997 (0.511)
527	141111111	0.188	-1.075 (0.404)	1.253 (0.897)	-1.041 (0.414)	1.477 (0.570)	-1.427 (0.338)
244	144111111	0.004	-0.774 (0.309)	1.541 (0.564)	-0.784 (0.314)	1.510 (0.578)	-1.128 (0.260)
74	333333333	0.032	1.255 (0.600)	-1.133 (0.655)	1.301 (0.577)	-1.263 (0.584)	0.314 (0.189)
90	223333222	0.559	-0.187 (0.190)	-0.526 (1.015)	-0.243 (0.203)	-1.329 (0.586)	-0.159 (0.165)

Note. TIMSS = Trends in International Mathematics and Science Study; SLM = Students Like Learning Mathematics; ERS = Extreme response style IRTree model; ORD = Ordinal IRTree model; Prob(Class2) = Estimated probability of belonging to ORD class; PSD = Posterior standard deviation.

The differences observed in the posterior means and standard deviations across models for the example patterns highlight some important differences between the mixture and single IRTree models. First, in comparing Respondents 4 and 603, the content trait estimates ($\hat{\theta}$) under the ERS model show no real difference between respondents, as the only information about the content trait is extracted from Node 1, and the two response patterns reflect identical outcomes for all items at Node 1 (i.e., $Y^*_{i,4,1} = Y^*_{i,603,1} = 0$ for all i). However, under both the ORD and mixture IRTree models, where Nodes 2 and 3 have the potential to contribute information to estimating θ, we can see differences such that Respondent 4 shows a lower content trait estimate than 603, as expected. A similar pattern of relationships is observed for Respondents 141 and 74. Thus, unlike the ERS model, the mixture model might be viewed as beneficial in allowing some information about the content trait to be extracted from the selection of extreme response categories.

Another aspect of using the mixture is seen when comparing the $\hat{\eta}$ for the mixture and ERS models. Note that the ERS estimates ($\hat{\eta}$) under the mixture model are pulled substantially closer to the prior mean (0) relative to the ERS model (except for Respondent 244 where the uncertainty of class membership is minimal), a consequence of the greater uncertainty regarding response style under the mixture model. Similarly, the posterior standard deviations of η are also seen to be larger under the mixture representation. These effects are particularly seen for Respondents 4 and 141. Note that while each of these respondents consistently selects the most extreme response option, the possibility under the mixture representation that such response patterns may also be due to extreme levels on the content trait makes the pattern less informative about extreme response style. As a result, extreme response style trait estimates

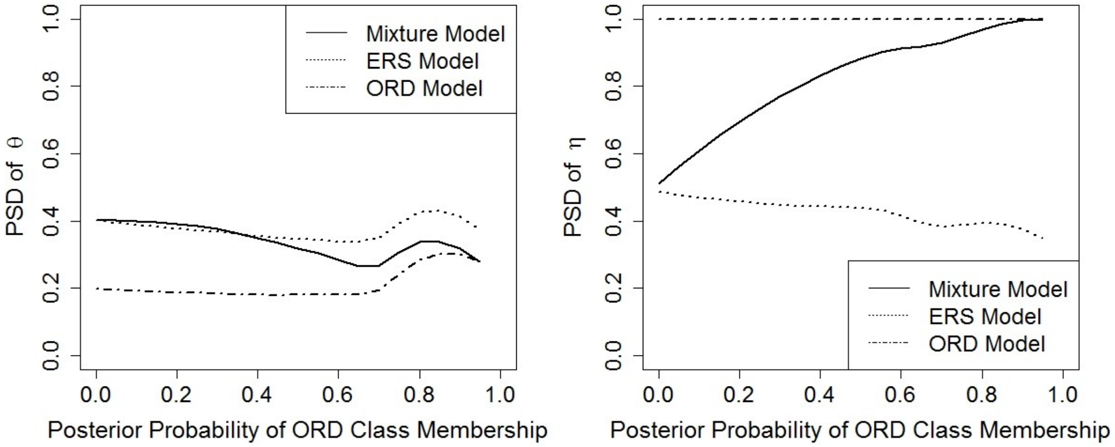
($\hat{\eta}$) are obtained under the mixture that are between the estimates under the single ERS IRTree model and the prior means (0). Such uncertainty of response process is also reflected in the content trait estimates ($\hat{\theta}$) under the mixture model that lie between the estimates under the single ERS and ORD IRTree models. Also, higher posterior standard deviations are observed under the mixture due to this uncertainty.

This change in posterior means and standard deviations of η shows how the mixture IRTree model takes into account the uncertainty of the respondent's class memberships. However, in certain instances, despite the presence of the mixture, class membership is quite clear, and this shrinkage effect is more minimal. An illustration is seen in comparing Respondents 527 and 244, each of whom select options from both extremes of the rating scale (both 1 and 4). For Respondent 244, who selects twice from the opposite end of the rating scale (4) relative to the modal response (1), the PSD of η does not get much larger in comparison to the ERS class, as the evidence for ERS is strong even in the application of a mixture model. For Respondent 527, who selects only once from the opposite end of the rating scale, the evidence is less strong, and the PSD of η is a little larger in comparison to that seen for the ERS class.

Figure 3.4 displays kernel-smoothed functions of the relationship between the posterior standard deviations (i.e., standard errors) of respondents' person parameters and the respondents' probability of belonging to the ORD class. The probability of ORD class membership indicates the uncertainty of the respondent's true class membership. Note that the curve for θ under the mixture model is quite consistently between those of the single tree ORD and ERS models, closer to the ERS model when the probability of being in the ORD class is low (see left panel). The mixture curve gets closer to the ORD model when the posterior probability of class membership is high. The curve lies between the two other curves (ERS and ORD) for the intermediate posterior probability of class membership values. Note that for η the PSD curve under the ORD model is drawn as a straight line at the value of 1, the standard deviation of the prior distribution of η. The mixture model produces progressively larger PSDs of η as the probability of belonging to ORD class increases, showing how the uncertainty of the response style estimates is substantially affected by the reduced certainty of membership in the ERS class. The single ERS IRTree model, in contrast, does not consider such uncertainty, and as a result reports η estimates with inflated levels of precision, as was observed in the simulation study.

Figure 3.4

Average Posterior Standard Deviations (PSD) of θ and η in Relation to the Posterior Probability of ORD Class Membership, for the Mixture and Single IRTree Models, for TIMSS SLM Scale Data



Note. TIMSS = Trends in International Mathematics and Science Study; SLM = Students Like Learning Mathematics; ERS = Extreme response style; ORD = Ordinal.

In summary, the pattern of results observed in the real data resembles quite closely the effects seen in the simulation. The assumption of a single class IRTree model, whether an ERS

IRTree or an ORD IRTree, overstates the precision of the estimated respondent traits when both response processes may be present in the respondent population. The use of a single IRTree model should thus be supported by evidence of its validity across all respondents; the results suggest that a mixture may likely be present, in which case model estimates should be sensitive to the unknown class to which the respondent belongs.

3.6 Conclusions

IRTree models have become a popular way of measuring response styles for self-report rating scale assessments. Such models associate a response process (represented in the form of a decision-making tree) with content and response style traits that underlie the different decision-making nodes within the tree. The separation of traits across nodes enhances the ability to measure response style traits but comes at the cost of assuming the same response process (associated with the same traits across respective steps in the process) for all respondents.

In this chapter, I demonstrate through a real data application the likelihood that no single IRTree may best characterize all respondents, and that a better representation of the response process may be achieved by allowing a mixture of trees. I demonstrate this possibility in the context of a commonly used IRTree model for extreme response style by considering also an alternative IRTree model that assumes the content trait is relevant at all decision nodes. The better comparative fit of the mixture model confirms such a mixture. Some of the more immediate implications of the mixture relate to the precision with which the content and response style traits are assumed to be measured. I suggest that the likely presence of a mixture introduces what should be viewed as the core challenge in attempts to measure response styles such as ERS, namely the uncertainty that exists as to whether the selection of an extreme response category is due to the content trait, a response style, or some combination of these factors. Through a mixture IRTree model approach, it becomes possible to see how the uncertainty of class membership impacts the estimation of both the content and response style traits. As expected, the extreme response styles tend to show larger posterior standard deviations (i.e., standard errors of estimates) when allowing for a mixture. I contend that this uncertainty, already an implicit part of both mixture IRT (von Davier & Rost, 1995) and multidimensional IRT (Bolt & Johnson, 2009) approaches to measuring response style, is important to consider when psychometric models are used to measure response styles. As Adams et al. (2019) note, attending to this uncertainty has various practical implications related to the design survey instruments, in particular, the value of having psychometrically heterogeneous items, as well as the relevance of having external criteria (e.g., anchoring vignettes, content-heterogeneous items) to more accurately measure response styles.

Although not explored in this chapter, a mixture tree representation arguably brings the IRTree methodology closer to that observed with mixture and multidimensional IRT models of response style. Future comparative studies of response style methods in this context would be useful. Along these lines, Meiser et al. (2019) also demonstrate the possibility that the outcome at a single decision tree node might be influenced simultaneously by both a content trait and a response style trait. In a similar way, such a model might also be anticipated to return reduced precision in the estimated response style trait due to the less certain roles the content and response style traits play in response category selection.

Of course, despite the use of a mixture model in this study, it is also conceivable that for an individual respondent the causes of extreme responses may also vary within a single respondent across items. For example, a respondent may for some items select an extreme category as a result of the content trait, and for other items a response style. This possibility was not considered in this article. It may be difficult to model such behavior unless a test is sufficiently long. Another issue not considered in this study concerns the potential for bias in the estimates of latent traits when failing to account for a true mixture. It is not difficult to envision scenarios whereby not only will precision be misestimated, but the trait estimates themselves become biased when a mixture is present but is not accounted for. I leave such investigation to future study.

In conclusion, I suggest that regardless of the methodology chosen in modeling response style, more attention should be devoted not just to focus on point estimates of content and response style traits returned but also the precision of those estimates. Such attention can make apparent where response styles can and cannot be successfully measured and will also make more apparent how different approaches to measuring response style may differ.

4 Evaluating Psychometric Differences Between Fast Versus Slow Responses on Rating Scale Items [3]

4.1 Introduction

Recent advances in technology have fostered the use of computer-based assessments in education and psychology. Unlike traditional paper-and-pencil tests, computer-based assessments enable the collection of *process data* in addition to item scores, potentially providing construct-relevant information about how individuals interact with items. Examples of process data include log data such as mouse clicks, timestamps, keystrokes, and action sequences (e.g., navigation between pages) as well as other types of information such as eye movement and brain imaging data. Because process data is naturally collected during assessments while not interfering with response process (respondents are nonreactive to the recording of process data), it provides opportunities for researchers to understand the underlying cognitive, behavioral, and psychological response process contributing to the final item responses.

The amount of time a respondent spends to solve or respond to an item is one of the most common forms of process data. Response time has been extensively studied especially in relation to cognitive measures, often with a goal to understand cognitive processes underlying response behavior (for an overview see De Boeck and Jeon, 2019). However, there has not been as much research examining the use of response times with noncognitive assessments (Henninger and Plieninger, 2021; Ranger, 2013). Oftentimes noncognitive assessments mea-

[3] This chapter is an adapted version of Kim and Bolt (2022) presented at the annual meeting of the National Council on Measurement in Education (NCME).

sure traits related to personality, behaviors, and attitudes by asking respondents to self-report to items using Likert-type rating scales. One concerning feature of this format is that the scales may be interpreted and used differently across respondents. This makes the item responses on rating scales often affected by *response styles* which refer to tendencies to over- or under-select specific response categories (Paulhus, 1991). Two of the most frequently observed response styles are extreme response style (ERS; the tendency to prefer extreme categories) and midpoint response style (MRS; the tendency to prefer a midpoint category).

Previous studies of rating scale measures imply that response times depend on the content trait. Ranger (2013) and Ranger and Ortner (2011) suggest that response times are generally shorter if the item response is more probable considering the content trait and psychometric characteristics of the item. For instance, on a rating scale from *extremely disagree* to *extremely agree*, a respondent with a high content trait level will be expected to select a category reflecting a high level of the trait (e.g., *extremely agree*) faster than a category that is unlikely (e.g., *extremely disagree*). This is largely in line with the findings in other studies demonstrating that fast responses are likely to occur when respondents have a strong self-schema relevant to the item content (Markus, 1977), which facilitates automated response processing (Akrami et al., 2007; Fazio et al., 1986; Holden et al., 1991). If a respondent has an accessible and strong trait relevant to item content, clear self-knowledge about the trait can lead to a high confidence and certainty about the response and, thus, the response can be given more easily and quickly (Arndt et al., 2018; Germeroth et al., 2015; Grant et al., 1994; Hanley, 1965). Henninger and Plieninger (2021), on the other hand, have recently illustrated that response times also depend on response styles. Their findings demonstrate that responses are faster when a respondent provides a response style-driven response. For example, respondents with a high ERS level generally respond faster when selecting extreme response categories and are slower for their nonextreme responses.

In this chapter, I explore the simultaneous relevance of the content trait and response styles on the association of item responses and response times in rating scales. Given that a primary purpose in modeling response times in relation to the content trait may be to understand the potential usefulness of response times toward improved trait estimation, it is important to understand how, if at all, response styles may interfere with such applications. I specifically look at how item response processes (attending to both content trait and response styles) differ across response times by examining psychometric differences across fast versus slow responses. Using several noncognitive assessment datasets, it is empirically illustrated whether and how fast versus slow responses may involve different item-person interactions. I specifically use residual item response times (after accounting for item and person effects) to classify each item-by-person response into a relatively fast or slow response class and inspect if there is any systematic difference in item properties across the two classes. The use of residual response times can provide insights into understanding the response behavior/process heterogeneity that may be present within respondents across response times.

The idea of this study parallels the approaches adopted in recent studies on cognitive assessments that have modeled local dependencies between item responses (i.e., response accuracy) and response times by allowing a heterogeneity in model parameters across response times (Bolsinova et al., 2017; DiTrapani et al., 2016; Molenaar & De Boeck, 2018; Partchev & De Boeck, 2012). These studies consistently found systematically varying item characteristics across slower and faster responses which indirectly speak to the different cognitive processes involved across response times. In a similar way, I intend to understand how response times may imply different things in rating scale assessments. It is expected that such an investigation can help to reveal the potential of response times in rating scale measurement for informing about respondents' response processes and ultimately improving the estimation of the content traits.

The remainder of this chapter is organized as follows. First, I describe an item response theory (IRT) model, specifically a multidimensional nominal response model (MNRM), that can attend to both content trait and response styles in rating scale assessments. I then demonstrate the way in which each item-by-person response is classified into a fast or slow response class and how the model is fitted to estimate separate item parameters for different response classes. Followed by the descriptions of data and how it was analyzed, I evaluate whether a separate estimation of item parameters can improve the model fit. I next provide an illustration of systematic differences in item parameter estimates across fast and slow responses, particularly focusing on item discrimination estimates on the content trait and response styles. I further demonstrate how the relative differences in discrimination estimates between fast versus slow responses vary across the content trait and response styles. Lastly, the results of this study are discussed providing insights into understanding response processes for slower and faster responses.

4.2 Multidimensional Nominal Response Model (MNRM) for Response Styles

In noncognitive self-report rating scale assessments, item responses are often reflective of both the intended-to-be-measured content trait and unintended-to-be-measured response styles such as extreme response style (ERS) and midpoint response style (MRS). One common and flexible item response theory (IRT) approach that can attend to both the content trait and response styles is a multidimensional nominal response model (MNRM; Bolt & Johnson, 2009; Falk & Cai, 2016; Falk & Ju, 2020; Thissen & Cai, 2016). The MNRM is a multidimensional extension of Bock (1972)'s nominal response model for polytomous items, introducing additional factor(s) to account for response style(s). The general form of MNRM (Thissen & Cai, 2016) can be specified as :

$$P(Y_{ji}=k|\theta_1,\theta_2,...,\theta_D)=\frac{\text{exp}(a_{i1}s_{k1}\theta_{j1}+a_{i2}s_{k2}\theta_{j2}+...+a_{iD}s_{kD}\theta_{jD}+c_{ik})}{\sum_{m=1}^{K}\text{exp}(a_{i1}s_{m1}\theta_{j1}+a_{i2}s_{m2}\theta_{j2}+...+a_{iD}s_{mD}\theta_{jD}+c_{im})} \tag{4.1}$$

where $P(Y_{ji}=k)$ is the probability that respondent $j(=1,2,...,J)$ selects category $k(=1,2,...,K)$ for item $i(=1,2,...,I)$. In this model, θ_{jd} represents respondent j's latent trait $d(=1,2,...,D)$; a_{id} is item i's discrimination parameter for latent trait θ_{jd}; and c_{ik} is a category-specific intercept parameter for category k of item i. The latent traits of the model are defined through scoring function parameter s_{kd} for each category k and latent trait θ_{jd}. The content trait and response styles can thus be defined in the model by assigning specific values to s_{kd}, as will be described shortly. In this study, for identification purposes, the first category intercept parameter for each item is fixed to 0 (i.e., $c_{i1} = 0$) and the latent factors are assumed to follow a multivariate normal distribution with the mean and variance of each factor set to 0 and 1, respectively.

The MNRM is a compensatory multidimensional IRT (MIRT) model that can assume item responses are influenced by both content trait and response styles in an additive way. As noted, each latent trait achieves its interpretation through its repective scoring function $\mathbf{s_d} = [s_{1d}, s_{2d}, ..., s_{Kd}]$ which represents the relationship between item categories and the latent trait. For instance, for an item measured on a 4-point Likert scale $(K = 4)$, where we can expect both content trait and extreme response styles (ERS) to affect the item responses $(D = 2)$, the two distinct dimensions can be defined through the functions $\mathbf{s_1} = [0, 1, 2, 3]$ and $\mathbf{s_2} = [1, 0, 0, 1]$, respectively. Note that the first scoring function implies high scores reflect high levels of the latent trait, while the second scoring function implies that extreme scores reflect higher levels of the latent trait. For item responses on rating scales with a middle category (e.g., 5-point rating scale items), midpoint response styles (MRS) can also be considered and defined by specifying a scoring function such as $\mathbf{s} = [0, 0, 1, 0, 0]$. The specification of scoring function $\mathbf{s_d}$ and corresponding latent traits is often largely based on substantive theory; in this case, I rely on the well-documented prevalence of ERS and MRS in rating scale instruments. Note that defining scoring functions distinctively across dimensions (i.e., less similar/redundant) can help disentangle content trait and response styles dimensions and lessen the difficulty in the estimation and interpretation of the latent dimensions (Falk & Cai, 2016; Falk & Ju, 2020). When these scoring functions are specified, we can also estimate item discrimination parameter a_{id} for each latent trait and item to quantify the overall effect of corresponding latent trait θ_d on item responses.

4.3 Distinguishing Fast Versus Slow Responses

To separately estimate item parameters for slow and fast responses, I first classify each *item response* into a fast or slow response class. Such a dichotomization of response times is based on the hypothesis that respondents would show different response behaviors depending on whether they responded fast or slow, adopting different response processing modes (e.g., automatic/spontaneous versus deliberate/controlled). To define item responses into fast or slow, I use residual response time in which item and person effects are removed. Through the use of *residual* response times, we can account for the effects of items and persons on response times and allow each respondent and item to have relatively fast and slow responses irrespective of the respondents' baseline response speed (e.g., reading speed, cognitive processing speed) and item attributes (e.g., item length, item complexity). It thus enables us to examine what response times within individuals imply, which can potentially lend insight into how individuals behave differently across items dependent on response times.

To this end, I consider an intercept-only cross-classified multilevel model treating each item response time as a value nested within items and respondents (having multiple memberships) and partial out item and respondent random effects. Specifically, the cross-classified model characterizes response times (RT) as

$$log(RT_{ji}) = \beta_0 + u_i + u_j + e_{ji} \tag{4.2}$$

$$u_i \sim N(0, \sigma^2_{u_1})$$

$$u_j \sim N(0, \sigma^2_{u_2})$$

$$e_{ji} \sim N(0, \sigma^2_e)$$

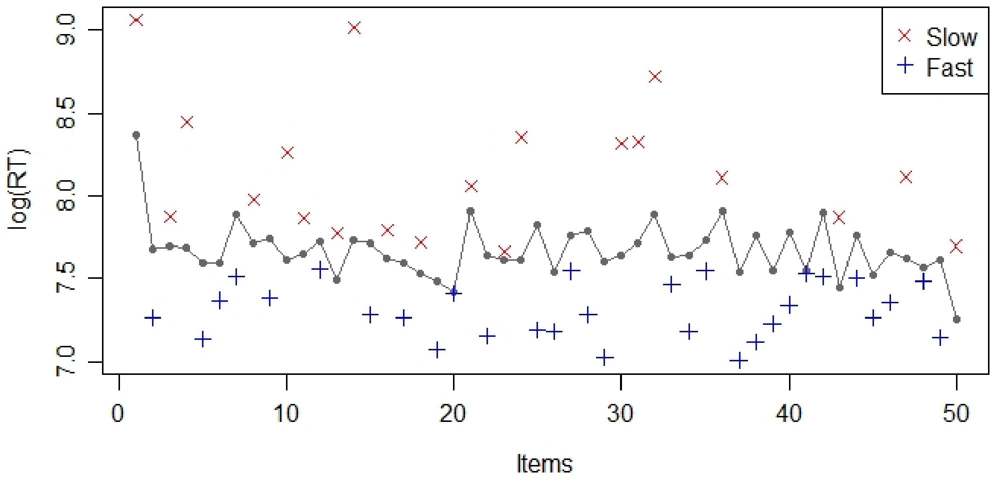
where β_0 is the model intercept, u_i and u_j each represents item (i) and respondent (j) random effects, and e_r denotes the residual response time for respondent j and item i. The response times are log-transformed to resolve the skewness of the response time data. I use the *lme4* package (Bates et al., 2015) in R (R Core Team, 2021) to estimate the model.

Using residual response time e_{ji} for each item response, I define each item-by-respondent response as fast or slow. It is classified as *fast* if e_{ji} is below 0 and *slow* if e_{ji} is above 0. Thus, as can be seen in Figure 4.1, responses are defined as *fast* (blue cross) if the response time is shorter than the expected response time based on the item intensity and person speed (filled circle), and *slow* (red cross) if the response time is longer than expected. The advantage of this approach is that it allows all respondents and items to provide data under fast and slow conditions, thus making it possible to link the separate calibrations of fast and slow responses.

Once the distinction of fast and slow item responses has been defined, the dataset can be represented as a mix of fast and slow response classes. To understand psychometric differences related to response times, item parameters for the two classes are separately estimated

Figure 4.1

An Example Illustration of Fitted (Filled Circles) and Observed (Blue or Red Cross Symbols) Log Item Response Time (RT) Within an Individual Across Items



by fitting the MNRM to data treating fast and slow responses as though they come from different items. Effectively, I merge each set of fast and slow responses as shown in Figure 4.2 and fit the model to the combined data. As can be observed from the figure, the resulting data has a doubled test length with exactly half of the item responses missing for each respondent, as each one of respondent's item responses is either in the fast or slow response type. Importantly, each respondent's latent traits are assumed invariant across response types so as to link the item parameter estimates across fast and slow response classes relying on a common respondent design.

Figure 4.2

An Example Illustration of Data Structure for a Combined Dataset of Fast and Slow Responses

	Slow responses					Fast responses				
Item Person	1	2	3	⋯	I	I+1	I+2	I+3	⋯	I+I
1	NA	NA	2	⋯	1	2	3	NA	⋯	NA
2	3	2	NA	⋯	NA	NA	NA	3	⋯	3
3	2	3	NA	⋯	3	NA	NA	1	⋯	NA
4	4	NA	NA	⋯	2	NA	4	4	⋯	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
J	1	2	NA	⋯	NA	NA	NA	3	⋯	4

Note. I = Number of items; J = Number of respondents. NA denotes missing values. The first I items (colored in red) are for slow responses and the latter I items (colored in blue) are for fast responses.

4.4 Methods

4.4.1 Data

I used three empirical datasets containing both item responses and response times to examine psychometric differences across slow and fast responses. Each dataset was accessed from *www.openpsychometrics.org*. The datasets result from three different online personality tests: the Fisher Temperament Inventory (FTI), the Big-Five Personality Test, and the Multidimensional Introversion-Extraversion Scales (MIES). Specific details of each test are described below.

Fisher Temperament Inventory (FTI) The Fisher Temperament Inventory (FTI) consists of 56 items measuring four broad temperament scales: Curious/Energetic, Cautious/Social

Norm Compliant, Analytical/Tough-minded, and Prosocial/Emphathetic. Sixteen items are used to measure each scale. The items are measured on a 4-point Likert rating scale (1 = Strongly Disagree, 2= Disagree, 3 = Agree, 4 = Strongly Agree) and were presented to each respondent in a random order. The accessed data was collected in 2019 containing item response and response time information (in milliseconds) from 4,967 respondents who had responded to an optional survey at the end of the assessment.

Big-Five Personality Test The Big-Five Personality Test measures five personality factors (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness) based on the Big-Five Factor Markers (Goldberg, 1992) from the International Personality Item Pool (IPIP). The test has 50 items in total (10 items for each dimension) and includes a number of reverse-worded items. Item responses are 5-point Likert ratings (1=Disagree, 3=Neutral, 5=Agree) and are presented in the same order to all respondents. The accessed data was collected from 2016 to 2018 and contains item responses and response times from 1,015,341 respondents who had agreed to provide their data for research use.

Multidimensional Introversion-Extraversion Scales (MIES) The Multidimensional Introversion -Extraversion Scales (MIES) test consists of 91 items measuring traits expected to show substantial differences between introverts and extraverts. The test includes reverse-worded items. Item responses are 5-point Likert ratings (1=Disagree, 2=Slightly disagree, 3=Neutral, 4=Slightly agree, 5=Agree) and are administered in a random order for each respondent. The accessed data contains responses from 7,188 respondents who had consented to provide their responses for research use.

Data preparation For each dataset, I initially carried out a two-step procedure for data cleaning. First, I removed the cases from the same IP address (indicating a duplicate submission or submissions from a shared network) for FTI and Big-Five Personality Test datasets, which had such information available. Second, item responses with extremely long or short response times were eliminated because such responses likely indicate invalid responses that may distort the results (Henninger & Plieninger, 2021; Mayerl, 2013). I treated item responses with log-response times that deviated more than ±2.5 standard deviation from the mean log-response times as outliers and removed them. After cleaning the data, I reverse-coded items that were reverse worded in the Big-Five Personality Test and MIES. To reduce the computational burden, I randomly sampled 2,000 respondents from each dataset for further analyses.

4.4.2 Analyses

The three datasets were each analyzed based on the approach described in Section 4.3 which I hereinafter refer to as *response time class analysis*. Note that, as already explained, the classes here are classes of item-by-person responses (not classes of items or persons) and are manifest classes (not latent classes). For model evaluation purposes, I also fitted the original MNRM as a baseline model, estimating a single set of item parameters for an item using all item responses. By comparing the two approaches, it can be evaluated whether allowing differences in item-person interaction across fast versus slow responses improves model fit.

As the three datasets had different numbers of content traits and response categories, the specifications of the MNRM fitted to each data were slightly different. For FTI data, the model incorporated four different content traits (corresponding to each of the four subscales) and ERS because four response categories are used in the test. The scoring functions were thus specified as $\mathbf{s_1} = \mathbf{s_2} = \mathbf{s_3} = \mathbf{s_4} = [0, 1, 2, 3]$ and $\mathbf{s_5} = [1, 0, 0, 1]$ for the content traits and ERS dimensions, respectively, and discrimination parameters for the content traits were constrained in a way that allows each item to only load on the content trait measured by the item. As to the Big-Five Personality Test and MIES data which involve five response categories, the content, ERS, and MRS traits were simultaneously considered. For the Big-Five Personality Test data, five different content traits corresponding to the five personality factors were assumed, while a single content trait was assumed for the MIES data. As a result, the scoring functions for the content, ERS, and MRS traits were respectively $\mathbf{s_1} = \mathbf{s_2} = \mathbf{s_3} = \mathbf{s_4} = \mathbf{s_5} = [0, 1, 2, 3, 4]$, $\mathbf{s_6} = [1, 0, 0, 0, 1]$, and $\mathbf{s_7} = [0, 0, 1, 0, 0]$ for the Big-Five Personality Test data and $\mathbf{s_1} = [0, 1, 2, 3, 4]$, $\mathbf{s_2} = [1, 0, 0, 0, 1]$, and $\mathbf{s_3} = [0, 0, 1, 0, 0]$ for the MIES data.

For model estimation, the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a, 2010b; Monroe & Cai, 2014; Yang & Cai, 2014), which can handle high dimensional estimation with greater efficiency, was used. I used *mirt* package (Chalmers, 2012) in R (R Core Team, 2021) for this purpose. As previously mentioned, the latent factors were assumed to follow a multivariate normal distribution with the mean and variance for each dimension fixed to 0 and 1, respectively. The correlations among the dimensions were allowed to be estimated. The intercept parameter for the first category was set to 0 across all items for model identification.

I first compared the model fit of under the response time class analysis to that of the baseline model which does not separately estimate item parameters across fast and slow responses. Two model fit indices were used for the comparisons: the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Confirming that a separate estimation of item parameters can improve the model fit, the estimated item parameters were compared across response classes for each dataset to consider any systematic differences in the estimates. Then, I further investigated how such differences varied across latent traits.

4.5 Results

The findings are presented in three subsections. I show the model fit comparison results, present systematic differences found in item parameter estimates across fast and slow response class, and further illustrate how these differences vary across the content and response styles traits. In this results section, I denote the content trait factor as θ, extreme response style trait factor as "ERS", and midpoint response style trait factor as "MRS" for convenience.

4.5.1 Model Fit

The estimated correlations among the content (θ) and response style traits (ERS and MRS) were first examined. For the FTI data, $\rho_{\theta,ERS}$ ranged from $-.159$ to $.132$ depending on the subscale; for BIG5 data, $\rho_{\theta,ERS}$ ranged from $-.234$ to $.024$, $\rho_{\theta,MRS}$ ranged from $-.073$ to $.112$, and $\rho_{ERS,MRS} = .052$; for MIES data, $\rho_{\theta,ERS} = .016$, $\rho_{\theta,MRS} = -.053$, and $\rho_{ERS,MRS} = 0.184$. Such low correlation estimates indicate that the response style dimensions appeared to be largely distinct from each other and the content traits.

Table 4.1 presents the model fit of the response time class analysis in comparison to the baseline model. As can be seen from the table, the response time class analysis produced lower AIC and BIC values compared to the baseline model for all three datasets, indicating a better model fit. Such results suggest that fast and slow responses are associated with different psychometric characteristics, a result seen previously with cognitive test instruments.

Table 4.1
Model Fit Comparisons Between the Response Time (RT) Class Analysis and the Baseline Model, for Fisher Temperament Inventory (FTI), Big-five Personality Test (BIG5), and Multidimensional Introversion-Extroversion Scales (MIES) Data

Data	Model	AIC	BIC
FTI	Baseline	231294.9	232919.2
	RT class analysis	226284.2	229476.6
BIG5	Baseline	246980.5	249058.4
	RT class analysis	244995.0	249033.3
MIES	Baseline	480088.0	483672.6
	RT class analysis	467343.1	474495.4

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

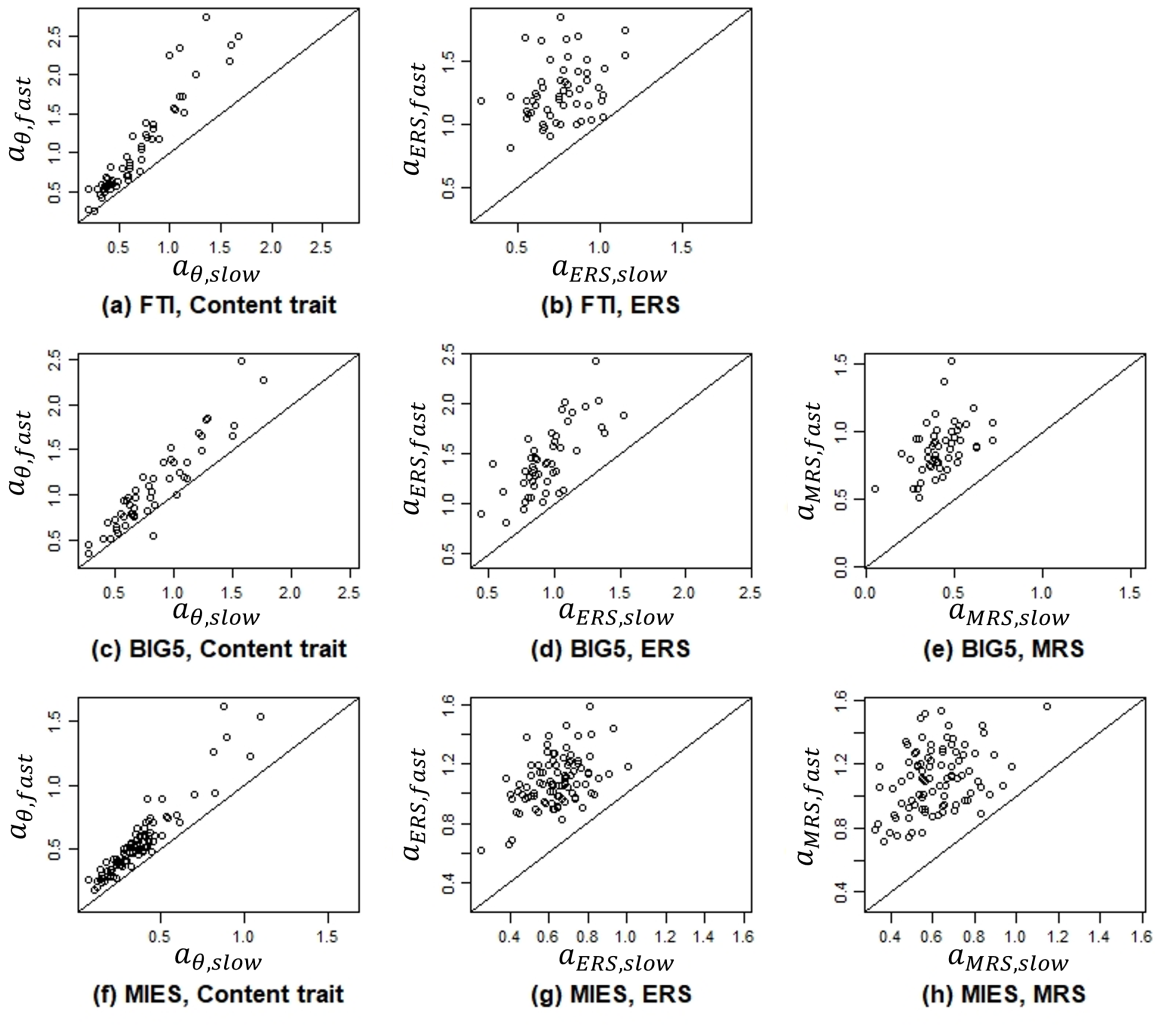
4.5.2 Differences in Item Parameter Estimates Across Fast and Slow Responses

Inspecting the differences in item parameter estimates across fast and slow response classes, a systematic difference was observed in item discrimination estimates on both the content trait and response styles. Figure 4.3 presents scatterplots of discrimination estimates derived from fast versus slow responses for each latent trait (i.e., content traits and response styles) for each dataset. Note that item discrimination estimates across fast and slow responses can be directly compared because the same respondents are present across the response classes, and respondent trait parameters are assumed to be invariant. We can see from the plots in

Figure 4.3 that the points consistently lie above the diagonal, indicating that discrimination estimates on content trait as well as response styles (ERS and MRS) are consistently higher for fast responses compared to slow responses.

Figure 4.3

Comparisons of the Estimated Item Discrimination Parameters on Content Trait (θ) and Response Styles (ERS and MRS) Across Fast and Slow Responses, for Fisher Temperament Inventory (FTI), Big-five Personality Test (BIG5) Personality Test, and Multidimensional Introversion-Extraversion Scales (MIES) Data



(a) FTI, Content trait

(b) FTI, ERS

(c) BIG5, Content trait

(d) BIG5, ERS

(e) BIG5, MRS

(f) MIES, Content trait

(g) MIES, ERS

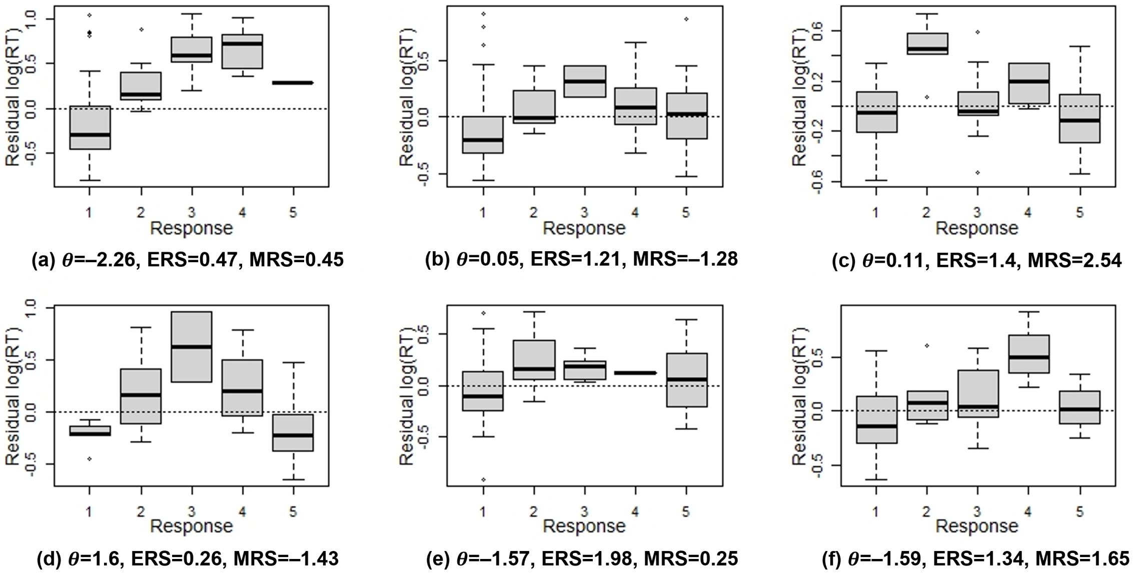
(h) MIES, MRS

Such a systematic difference in discrimination estimates across slow and fast responses implies that faster responses are simultaneously more reflective of both the content trait and response styles than slower responses. To examine this in more detail, I looked at response time distributions of responses at individual level. Figure 4.4 provides an illustration of some example distributions of residual response times for each response category within individuals with different levels of content trait and response styles. As expected, it is observed that responses matching to the levels of the content trait and response styles are generally given faster than other responses. For example, respondents with high or low content trait level tended to give faster responses on extreme response categories in the direction corresponding to their trait level (as can be seen in plots a, d, e, and f). As to the response styles, respondents with high ERS levels generally selected extreme categories faster than nonextreme categories (see plots b, c, e, and f) and those with high and low MRS levels tended to select the midpoint category relatively faster (c and f) and slower (b and d), respectively. One important observation is that the plots exhibit simultaneous effects of the content trait and response styles on response times. Respondents shown in plots (b) and (c) illustrate simultaneous effects of the two response styles (ERS and MRS) on response times, while plots (d), (e), and (f) show simultaneous effects of the content trait and response styles. All in all, the plots in Figure 4.4 consistently demonstrate that faster responses (under the dotted horizontal line at 0) do better reflect the respondents' content trait and response styles levels, as implied from the higher item discrimination estimates for faster responses.

In addition to the discrimination parameters, I was also interested in looking at differences in category intercept estimates across fast versus slow responses. For this purpose, I

Figure 4.4

Example Distributions of Residual Log Response Time (RT) for Each Response Category Selection Within Individuals Selected from Multidimensional Introversion-Extraversion Scales (MIES) Data



(a) θ=–2.26, ERS=0.47, MRS=0.45

(b) θ=0.05, ERS=1.21, MRS=–1.28

(c) θ=0.11, ERS=1.4, MRS=2.54

(d) θ=1.6, ERS=0.26, MRS=–1.43

(e) θ=–1.57, ERS=1.98, MRS=0.25

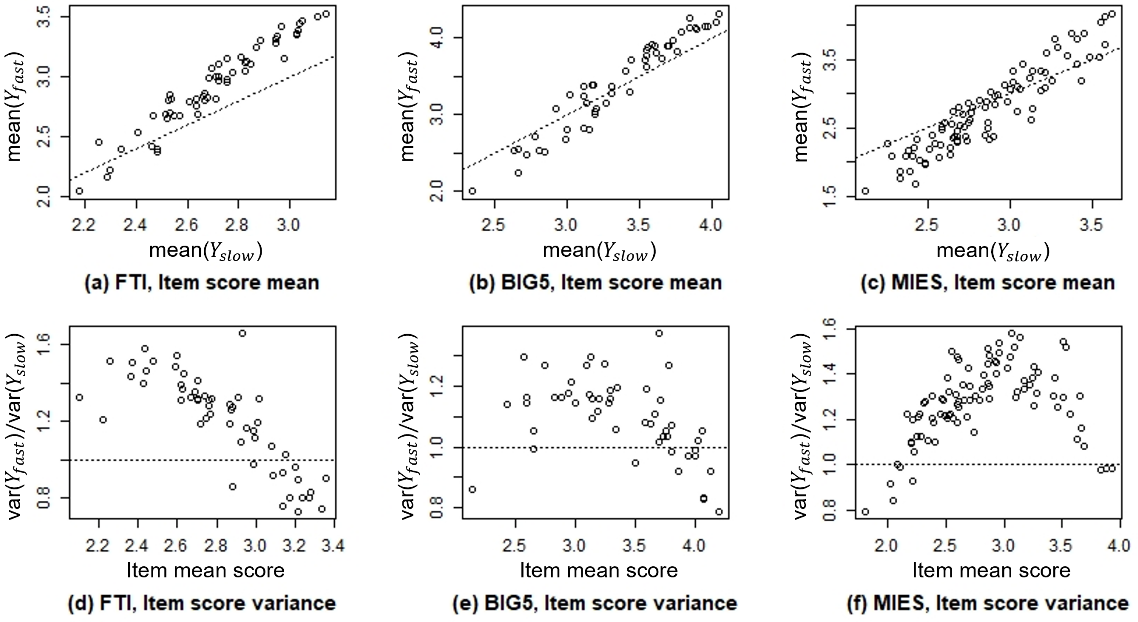
(f) θ=–1.59, ERS=1.34, MRS=1.65

Note. θ = Content Trait; ERS = Extreme Response Style; MRS = Midpoint Response Style.

compared the mean and variance of item scores for slow and fast responses, which are easier to interpret while providing similar information to that of the category intercepts. As shown in the scatterplots of item mean scores for fast versus slow responses in Figure 4.5 (plots a, b, and c), fast responses tended to show lower mean scores when items were more difficult (i.e., items more difficult to agree with), but higher mean scores as items became easier. As to the variance of item responses, it appeared that the variance of fast responses relative to that of slow responses was generally larger for items with overall mean scores (across slow and fast responses) near the center of the scale (i.e., 2.5 for FTI and 3 for BIG5 and MIES data), and it reduced as the mean score either increased or decreased (see plots d, e, and f in Figure 4.5). These results consistently suggest that faster responses likely involve more extreme responses than slower responses, a result consistent with previous studies illustrating the association between fast response times and extreme ratings (Casey & Tryon, 2001; Grant et al., 1994; Kuiper, 1981).

Figure 4.5

Comparisons of Item Mean Scores Across Fast and Slow Responses For Each Dataset (upper) and Ratio of Item Score Variance for Fast and Slow Responses Plotted Against Item Mean Scores For Each Dataset (lower)



(a) FTI, Item score mean **(b) BIG5, Item score mean** **(c) MIES, Item score mean**

(d) FTI, Item score variance **(e) BIG5, Item score variance** **(f) MIES, Item score variance**

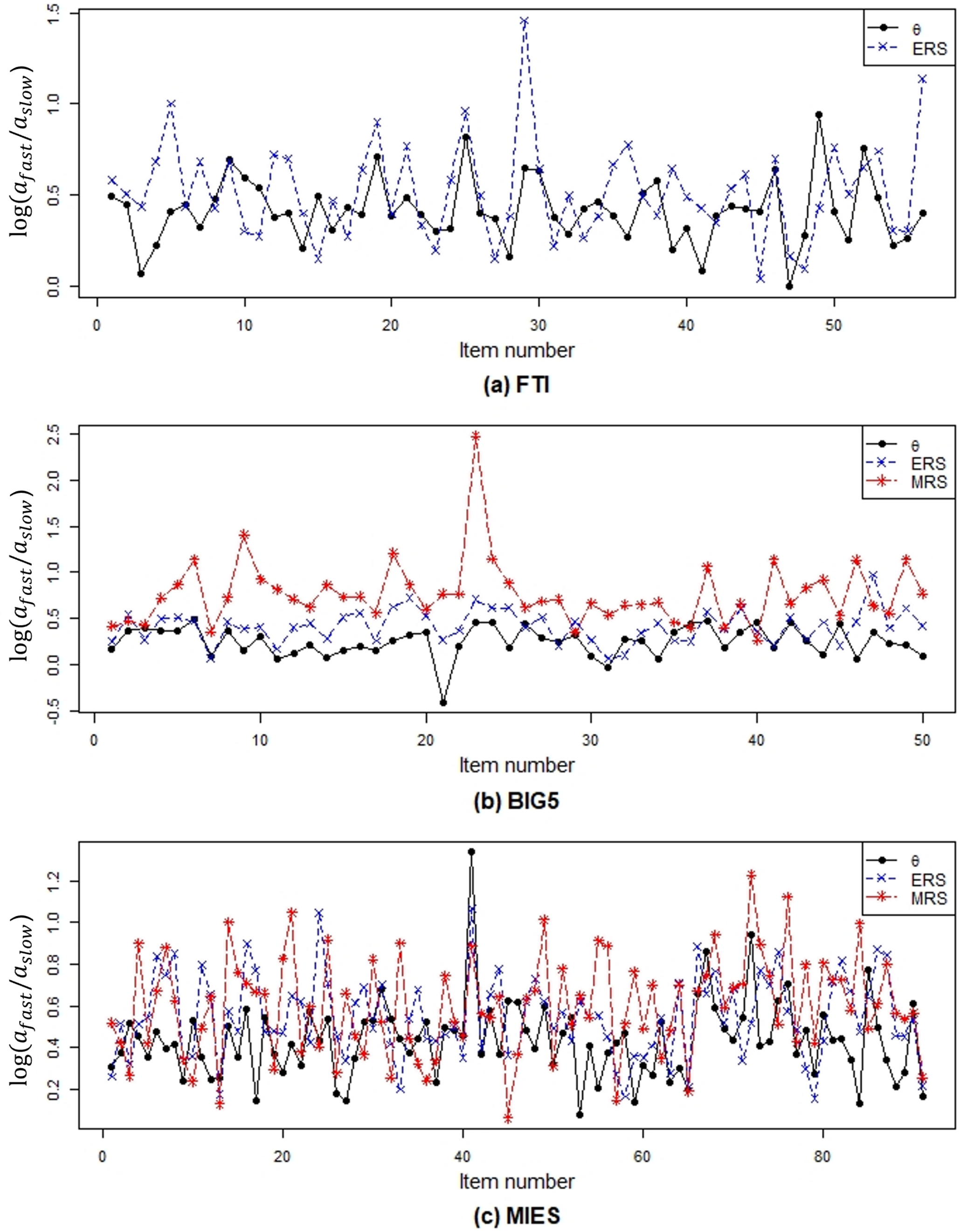
4.5.3 Relative Differences in Discriminations Across Fast and Slow Responses

I further compared relative differences in discrimination estimates for fast versus slow responses across the content and response style traits. Specifically, the log of the ratio of discrimination estimates from fast to slow responses ($= log(a_{fast}/a_{slow})$) on each latent trait was obtained and compared to examine which latent factor shows a greater difference in the influence on item responses across fast and slow responses. Figure 4.6 demonstrates how the relative differences in item discriminations between fast and slow responses differ across the content trait (θ), ERS, and MRS for each dataset. As discussed above, for most items, $log(a_{fast}/a_{slow})$ values are above 1 for all latent traits in all three datasets indicating that slow responses are less affected by both the content trait and response styles than fast responses. The values, however, generally seem to be larger for response styles traits (i.e., ERS and MRS) compared to the content trait (θ). This indicates that the relative amount of information that fast responses contain compared to slow responses is greater for response styles than content trait implying that respondents tend to behave differently across fast and slow responses more in relation to response styles than to the content trait.

Moreover, between the two response styles, the MRS tended to have a larger difference in fast and slow discrimination estimates compared to the ERS. In other words, the relative amount of information fast responses possess compared to slow responses appeared to be generally greater for MRS. This is possibly because fast midpoint responses are mainly reflective of MRS whereas fast extreme responses may often contain mixed information about the content trait and ERS. Specifically, responses on a midpoint category usually take longer as they are cognitively more demanding and are associated with unclear, ambiguous items (Kulas & Stachowski, 2009). It is highly probable that fast responses on a midpoint category are due to a high level of MRS (Henninger & Plieninger, 2021) providing less contaminated information on MRS. On the other hand, fast responses on extreme categories can be given either due to extreme levels of content trait or/and a high level of ERS, which makes it difficult to separate/distinguish the mixed effects of the two traits on fast extreme responses.

Figure 4.6

Comparisons of the Relative Discrimination Estimates for Fast Versus Slow Responses in Relation to Content Trait (θ) and Response Styles (ERS and MRS), for Fisher Temperament Inventory (FTI), BIG5 Personality Scale, and Multidimensional Introversion-Extraversion Scales (MIES) Data.



(a) FTI

(b) BIG5

(c) MIES

4.6 Conclusion and Discussion

Due to the increased use of computer-based assessments, response times are now a common and accessible form of process data that can provide useful information for understanding underlying response processes or behaviors. As there has not yet been much research investigating response times with rating scale measures, this chapter explored the association between item responses and response times in noncognitive assessments by attending to two types of traits that have been found relevant in previous studies. Ranger (2013) and Ranger and Ortner (2011) showed how item response times are influenced by the content trait level, while Henninger and Plieninger (2021) demonstrated the impact of response styles on item response times.

This study emphasizes the simultaneous effects of the content trait and response styles on response times in rating scale measurement through real data applications. Following parallel work with cognitive assessments, I examined psychometric differences between fast and slow responses and showed how items display different psychometric properties contingent on whether the item response is faster or slower than expected. In particular, systematically higher item discriminations are seen for faster responses on both the content trait and response styles, indicating that faster item responses are generally more influenced by the content trait as well as response styles. Further, response styles appear to have a stronger influence on response times than the content traits, as evidenced by the greater difference in discriminations between fast and slow responses for the response style traits. Such findings imply that respondents tend to behave differently across items depending on whether they responded slower or faster, especially in relation to response styles.

These findings are consistent with previous studies that have examined response times with rating scale measures. According to Callegaro et al. (2009) and Henninger and Plieninger (2021), fast responses on rating scale items may be due to a respondent's (1) high confidence in the item response and/or (2) shallower cognitive process. Many studies have reported that respondents tend to give fast and automated responses when they have a strong self-schema related to the item (Akrami et al., 2007; Fazio et al., 1986; Grant et al., 1994; Hanley, 1965; Holden et al., 1991). In contrast, slow responses are given when respondents are uncertain about the responses or provide "controlled" responses (e.g., editing or fake responses; Holden, 1995; Holden & Hibbs, 1995; Holden et al., 1992; Holtgraves, 2004; Monaro et al., 2021; Roma et al., 2018). Fast responses may thus be associated with higher confidence/certainty, extreme item responses, and higher consistency/stability (Arndt et al., 2018; Casey & Tryon, 2001; Germeroth et al., 2015; Grant et al., 1994; Kuiper, 1981). At the same time, fast and spontaneous responses are also known to occur when respondents go through a shallow and superficial cognitive process (Callegaro et al., 2009) which may arise as a response styles-driven response process. Along these lines, Henninger and Plieninger (2021) illustrated that response times are shorter if a respondent makes a response-style-driven category selection. The results of this study are consistent with these prior studies in showing that fast responses are likely to be highly reflective of both the content trait as well as response styles.

While prior studies either focused on the content trait or response styles in exploring response times with rating scale measures, this study attended to the simultaneous influence of both traits. One practical implication is that response styles will likely need to be attended to if one wants to use response times to improve the estimation of the content trait. In cognitive tests, it might be easy to use response times for improving the estimation of ability traits. In noncognitive assessments, however, this study's results indicate that the informative value of response times for the estimation of construct measures might be relatively limited given the interference from response styles, despite the previous findings illustrating that response times can inform about the construct measures (Ranger, 2013; Ranger & Ortner, 2011).

The results highlight another place in which the effects of the content trait and response styles in item response processes may be confounded (Adams et al., 2019). Item response tree (IRTree) models for response styles have received much attention for modeling response styles in rating scale items. The models typically assume that the effects of the content trait and response styles can be separately modeled by attaching each trait to a different subprocesses of item responses. From another perspective, however, the content trait and response styles may be inextricably confounded, possibly due to the presence of a mixture of respondent types (Kim & Bolt, 2021). In this study, the same form of confounding seems to be present in response times. Specifically, it is difficult to distinguish whether fast extreme responses are given due to the extreme level of the content trait or extreme response styles. As a consequence, it does not seem that attending to response times necessarily provides panacea to the confounding of these traits in the selection of the item response.

Overall, this study lends insight into understanding response processes across response times with regard to the content trait and response styles. Nevertheless, there remain additional directions for research. I primarily focused on examining psychometric differences between fast and slow responses by comparing item properties. Given that the ultimate goal of modeling response time is to improve the estimation of content traits, future work may further examine and evaluate whether the use of response times can actually improve the trait estimation. Such an investigation can be particularly useful and interesting to look at with a shorter test where the estimation of person traits solely based on item responses can be relatively less accurate. In addition, while this study mainly focused on two types of response styles (i.e., ERS and MRS), researchers may also consider other types of response styles such as acquiescent response style (ARS; the tendency to agree irrespective of item content) or socially desirable responding (the tendency to respond in a way to look good). Moreover, though I have eliminated outlier item responses with extremely fast or slow response times from the analyses, such outlier response times may be useful in informing about careless responding behaviors which can be interesting to look at.

In this study, item responses were classified into a fast or slow response class based on residual item response times in which the effects of items and persons on response times were removed. The use of residual response times made it available to examine whether and how respondents behave differently across items depending on whether they responded relatively faster or slower than is expected, which further helped to understand how response times are associated with item responses through the content trait and response styles within individuals. However, future studies may use response times instead of the residuals and explore the effects of item features (e.g., item position, item length/complexity) on response times to facilitate a better understanding of item-person interactions. Moreover, as decisions for distinguishing fast and slow responses are arbitrary, future studies may consider applying different split methods or criteria to classify response times which may yield different results and implications. For instance, researchers may use median values of response times to distinguish fast and slow responses, or even quartiles to split response times into four classes instead of two as in this study. A stochastic way of finding a cut-off value that best explains the data can also be considered as was done in Molenaar and De Boeck (2018). Furthermore, rather than dichotomizing response times that is likely to reduce the amount of information, future work may use response times as a continuous variable and examine the psychometric heterogeneity across response times by explicitly modeling the effects of response times on item parameters (Bolsinova et al., 2017).

We should also be cautious in generalizing the results of this study as different findings can be observed with different populations or tests. For instance, the three datasets used in this study are collected from a freely accessible website, which usually do not require a high motivation or cognitive efforts of respondents. It is hence plausible that using data from other types of tests collected in different ways (particularly in a way that can draw highly motivated respondents) can lead to different findings. Lastly, it would be interesting and useful to use other types of process data such as eye tracking or brain imaging data, which are becoming more available, to validate the findings from response times or to further improve the understanding of item response process in rating scale measurement.

5 Conclusions and Future Directions

5.1 Conclusions

Throughout this dissertation, the three studies coherently illustrated that understanding and incorporating item response process into psychometric models can contribute to improving the measurement of latent constructs and obtaining richer information about respondents and items/tests. In Chapter 2, the proposed noncompensatory MIRT model appeared to provide a more intuitive and practical approach for modeling passage-based test item responses with less model complexity compared to the bifactor model. The model successfully captured local item dependencies present in the passage-based structure while providing a better representation of the underlying cognitive process/components for solving passage-based items. The parameterization of the model also seemed to provide a more intuitive and easier interpretation of item properties, potentially improving our understanding of how items function with respect to the underlying components/processes.

In Chapter 3, I demonstrated that the mixture IRTree model can produce more valid and accurate measurement results in rating scale assessments, by attending to the heterogeneity of item response process across respondents as well as the uncertainty of item response process within respondents. The study specifically showed the capacity of the mixture model to account for the response process uncertainty of an individual by reflecting it to the precision of latent trait estimates, and highlighted the potential biases when ignoring such an uncertainty; namely, we can overstate the precision with which latent constructs are measured.

Chapter 4 suggested that using response time information can lend insight into understanding item response process heterogeneity in rating scale measurement. The results illustrated that both the content trait and response styles had differential effects on item responses dependent on item response times; specifically, faster item responses were more influenced by the content and response style traits compared to slower responses. Such differences imply a response process heterogeneity within individuals and that the content trait and response styles may have simultaneous effects in item response process. It is emphasized that the effects of the content and response style traits may be confounded, which may to some extent limit the use of response time information in rating scale measurement.

In conclusion, the measurement models presented in the three chapters contribute to improving our understanding of underlying item response processes and behaviors in cognitive and noncognitive assessments. Also, these studies illustrate the potentials and benefits of incorporating item response process into psychometric modeling, providing a basis for future research.

5.2 Future research directions

It is undeniable that psychometrics is rapidly evolving, bringing many new opportunities and challenges to researchers. For instance, due to the increased use of computer-based testing, innovative assessments and novel item formats have been developed and new forms of data (i.e., process data) such as response times and action sequences have become available. There is also an increasing interest in integrating learning and assessment, and measurement goals are expanding from assessment of learning to assessment for learning. Such advancement and expansion of the area increases the need for understanding and modeling item response process, through which we can enhance the validation of novel item and test formats, obtain more informative and diagnostic information from assessments, improve the measurement of latent constructs, and maximize the use of data available for understanding respondents and items.

In this sense, I suggest two broad directions, which can interplay with each other, for future research on item response process: 1) to explore the use of process data in cognitive and noncognitive assessments to enhance the understanding of response processes, and 2) to investigate item response processes in relation to innovative assessments and novel item types. As shown in Chapter 4 of this dissertation, the use of process data, such as response times, can improve our understanding of underlying cognitive or psychological response processes that we cannot explicitly observe from final item responses (De Boeck & Jeon, 2019). I believe making use of process data can better the measurement of constructs and lend deeper insight into student development and learning. Future studies may thus investigate how response times and other types of process data (e.g., page navigations and mouse clicks) can help to detect and understand the response process heterogeneity in cognitive and noncognitive assessments. Specifically, we may explore the usefulness of various forms of process data and develop psychometric models incorporating process data information to better estimate and distinguish traits intended to be measured and not intended to be measured (e.g., response styles). Future work may also consider using machine learning techniques to classify or predict response behaviors of individuals, broadening the scope of research on response process heterogeneity.

Moreover, we may further investigate response behaviors and cognitive strategies for solving or responding to test items, especially in relation to new formats of assessments and items. For instance, future work may include a further exploration of the noncompensatory approach illustrated in Chapter 2 in relation to innovative assessments. As demonstrated in the chapter, the noncompensatory model may provide a useful framework to understanding and modeling problem-solving processes in new formats of assessments, such as game-based or simulation-based assessments, because such tests often measure integrated competencies requiring multiple components to process. Thus, it would be interesting to look at how the noncompensatory model can be adapted to novel types of assessments measuring complex and integrated abilities (e.g., collaborative problem-solving skills, creating thinking). Further, it may be even more intriguing to incorporate process data (including eye movement and brain imaging data which can be a potent source of information for understanding item response process) into the model to potentially provide an improved understanding of individuals' item solving/response process in assessments.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*(4), 311–329. https://doi.org/10.1177/014662169602000402

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 131–154. https://doi.org/10.1111/bmsp.12169

Akrami, N., Hedlund, L.-E., & Ekehammar, B. (2007). Personality scale response latencies as self-schema indicators: The inverted-u effect revisited. *Personality and Individual Differences*, *43*(3), 611–618. https://doi.org/https://doi.org/10.1016/j.paid.2006.12.005

Arndt, C., Lischetzke, T., Crayen, C., & Eid, M. (2018). The assessment of emotional clarity via response times to emotion items: Shedding light on the response process and its relation to emotion regulation strategies. *Cognition and Emotion*, *32*(3), 530–548. https://doi.org/10.1080/02699931.2017.1322039

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/https://doi.org/10.1016/j.paid.2005.10.018

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

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

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

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*, 1126–1148. https://doi.org/https://doi.org/10.1007/s11336-016-9537-6

Bolt, D. M. (2019). Bifactor MIRT as an appealing and related alternative to CDMs in the presence of skill attribute continuity. In M. von Davier & Y. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 395–417). Springer Cham. https://doi.org/10.1007/978-3-030-05584-4

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335–352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395–414. https://doi.org/10.1177/0146621603258350

Bolt, D. M., & Liao, X. (2022). Item complexity: A neglected psychometric feature of test items? *Psychometrika*. https://doi.org/10.1007/s11336-022-09842-0

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. https://doi.org/10.1177/2167702616657069

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168. https://doi.org/https://doi.org/10.1007/BF02294533

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248. https://doi.org/https://doi.org/10.1037/a0023350

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T.-Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *103*(1), 5–25. https://doi.org/10.1177/075910630910300103

Casey, M. M., & Tryon, W. W. (2001). Validating a double-press method for computer administration of personality inventory items. *Psychological Assessment*, *13*(4), 521–530. https://doi.org/10.1037//1040-3590.13.4.521

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, C., Lee, S.-y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among east asian and north american students. *Psychological Science*, *6*(3), 170–175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*, 102. https://doi.org/10.3389/fpsyg.2019.00102

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

de Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(1), 104–115. https://doi.org/10.1509/jmkr.45.1.104

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145–168. https://doi.org/https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2016). Partially compensatory multidimensional item response theory models: Two alternate model forms. *Educational and Psychological Measurement*, *76*(2), 231–257. https://doi.org/10.1177/0013164415589595

DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence*, *56*, 82–92. https://doi.org/https://doi.org/10.1016/j.intell.2016.02.012

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138–149. https://doi.org/https://doi.org/10.1037/met0000121

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186. https://doi.org/https://doi.org/10.1007/BF02294171

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. https://doi.org/10.1037/met0000059.supp

Falk, C. F., & Ju, U. (2020). Estimation of response styles using the multidimensional nominal response model: A tutorial and comparison with sum scores. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00072

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037//0022-3514.50.2.229

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Germeroth, L. J., Wray, J. M., & Tiffany, S. T. (2015). Response time to craving-item ratings as an implicit measure of craving-related processes. *Clinical psychological science : a journal of the Association for Psychological Science*, *3*(4), 530–544. https://doi.org/10.1177/2167702614542847

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436. https://doi.org/https://doi.org/10.1007/BF02295430

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Grant, M. J., Button, C. M., & Noseworthy, J. (1994). Predicting attitude stability. *Canadian Journal of Behavioural Science*, *26*(1), 68–84. https://doi.org/10.1037/0008-400X.26.1.68

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328–351. https://doi.org/10.1086/269326

Hanley, C. (1965). Personality item difficulty and acquiescence. *Journal of Applied Psychology*, *49*(3), 205–208. https://doi.org/10.1037/h0022107

Henninger, M., & Plieninger, H. (2021). Different styles, different times: How response times can inform our knowledge about the response process in rating scale measurement. *Assessment*, *28*(5), 1301–1319. https://doi.org/10.1177/1073191119900003

Holden, R. R. (1995). Response latency detection of fakers on personnel tests. *Canadian Journal of Behavioural Science*, *27*(3), 343–355. https://doi.org/10.1037/0008-400X.27.3.343

Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment*, *3*, 111–118. https://doi.org/10.1037/1040-3590.3.1.111

Holden, R. R., & Hibbs, N. (1995). Incremental validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality*, *29*(3), 362–372. https://doi.org/https://doi.org/10.1006/jrpe.1995.1021

Holden, R. R., Kroner, D. G., Fekken, C. G., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology*, *63*(2), 272–279. https://doi.org/10.1037/0022-3514.63.2.272

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172. https://doi.org/10.1177/0146167203259930

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, *2*, 41–54. https://doi.org/https://doi.org/10.1007/BF02287965

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*(3), 261–277. https://doi.org/https://doi.org/10.1037/1082-989X.2.3.261

Ip, E. H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*, 73–91. https://doi.org/https://doi.org/10.1007/BF02294187

Ip, E. H., Molenberghs, G., Chen, S.-H., Goegebeur, Y., & De Boeck, P. (2013). Functionally unidimensional item response models for multivariate binary data. *Multivariate Behavioral Research*, *48*(4), 534–562. https://doi.org/10.1080/00273171.2013.796281

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, *67*(3), 367–386. https://doi.org/10.1007/BF02294990

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 395–416. https://doi.org/https://doi.org/10.1348/000711009X466835

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

Jiao, H., Lissitz, R. W., & Zhan, P. (2017). A noncompensatory testlet model for calibrating innovative items embedded in multiple contexts. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 117–138). Information Age Publishing.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*(2), 264–277. https://doi.org/10.1177/0022022104272905

Kellner, K. (2019). *jagsUI: A wrapper arouns 'rjags' to Streamline 'JAGS' analyses*. https://github.com/kenkellner/jagsUI

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, *49*(2), 161–177. https://doi.org/10.1080/00273171.2013.866536

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

Kim, N., & Bolt, D. M. (2022). *Understanding response process heterogeneity across fast and slow responses in noncognitive tests* [Paper presentation]. The annual meeting of the National Council on Measurement in Education (NCME), San Diego, CA, United States.

Kim, N., Bolt, D. M., & Wollack, J. (2022). Noncompensatory mirt for passage-based tests. *Psychometrika*. https://doi.org/10.1007/s11336-021-09826-6

Kuiper, N. A. (1981). Convergent evidence for the self as a prototype: The "inverted-u rt effect" for self and other judgments. *Personality and Social Psychology Bulletin*, *7*(3), 438–443. https://doi.org/10.1177/014616728173012

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*(3), 489–493. https://doi.org/https://doi.org/10.1016/j.jrp.2008.12.005

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21. https://doi.org/10.1177/0146621605275414

Luecht, R., & Ackerman, T. A. (2018). A technical note on irt simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and Practice*, *37*(3), 65–76. https://doi.org/https://doi.org/10.1111/emip.12185

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523–547. https://doi.org/https://doi.org/10.1007/BF02294327

Markus, H. R. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, *35*, 63–78.

Mayerl, J. (2013). Response latency measurement in surveys. detecting strong attitudes and response effects. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2013-00005

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*(7), 1539–1550. https://doi.org/https://doi.org/10.1016/j.paid.2008.01.010

Meiser, T., Plieninger, H., & Henninger, M. (2019). Irtree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 501–516. https://doi.org/https://doi.org/10.1111/bmsp.12158

Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279–297. https://doi.org/10.1007/s11336-017-9602-9

Monaro, M., Mazza, C., Colasanti, M., Ferracuti, S., Orr, G., di Domenico, A., Sartori, G., & Roma, P. (2021). Detecting faking-good response style in personality questionnaires with four choice alternatives. *Psychological Research*, *85*, 3094–3107. https://doi.org/10.1007/s00426-020-01473-3

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis–Hastings Robbins–Monro algorithm. *Educational and Psychological Measurement*, *74*(2), 343–369. https://doi.org/10.1177/0013164413499344

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261–286. https://doi.org/https://doi.org/10.1111/j.1467-6494.2008.00545.x

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

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32. https://doi.org/https://doi.org/10.1016/j.intell.2011.11.002

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* (pp. 17–59). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

Plummer, M. (2017). *JAGS version 4.3.0 user manual [computer software manual]*. https://sourceforge.net/projects/mcmc-jags

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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

Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological test and assessment modeling*, *55*, 361.

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*(2), 389–406. https://doi.org/10.1177/0013164410382895

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/https://doi.org/10.1007/978-0-387-89976-3

Rijmen, F. (2009). Efficient full information maximum likelihood estimation for multidimensional IRT models. *ETS Research Report Series*, *2009*(1), i–31. https://doi.org/https://doi.org/10.1002/j.2333-8504.2009.tb02160.x

Roma, P., Verrocchio, M. C., Mazza, C., Marchetti, D., Burla, F., Cinti, M. E., & Ferracuti, S. (2018). Could time detect a faking-good attitude? a study with the MMPI-2-RF. *Frontiers in psychology*, *9*, 1064. https://doi.org/10.3389/fpsyg.2018.01064

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. https://doi.org/https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 51–73). Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/9781315119144

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260. https://doi.org/https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Van Vaerenbergh, Y., & Thomas, T. D. (2012). 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 von Davier, M., & Rost, J. (1995). Polytomous mixed rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). Springer New York. https://doi.org/10.1007/978-1-4612-4230-7_20

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. https://doi.org/https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? what is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22–29. https://doi.org/https://doi.org/10.1111/j.1745-3992.1996.tb00803.x

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, 479–494. https://doi.org/https://doi.org/10.1007/BF02293610

Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis–Hastings Robbins–Monro algorithm. *Journal of Educational and Behavioral Statistics*, *39*(6), 550–582. https://doi.org/10.3102/1076998614559972

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

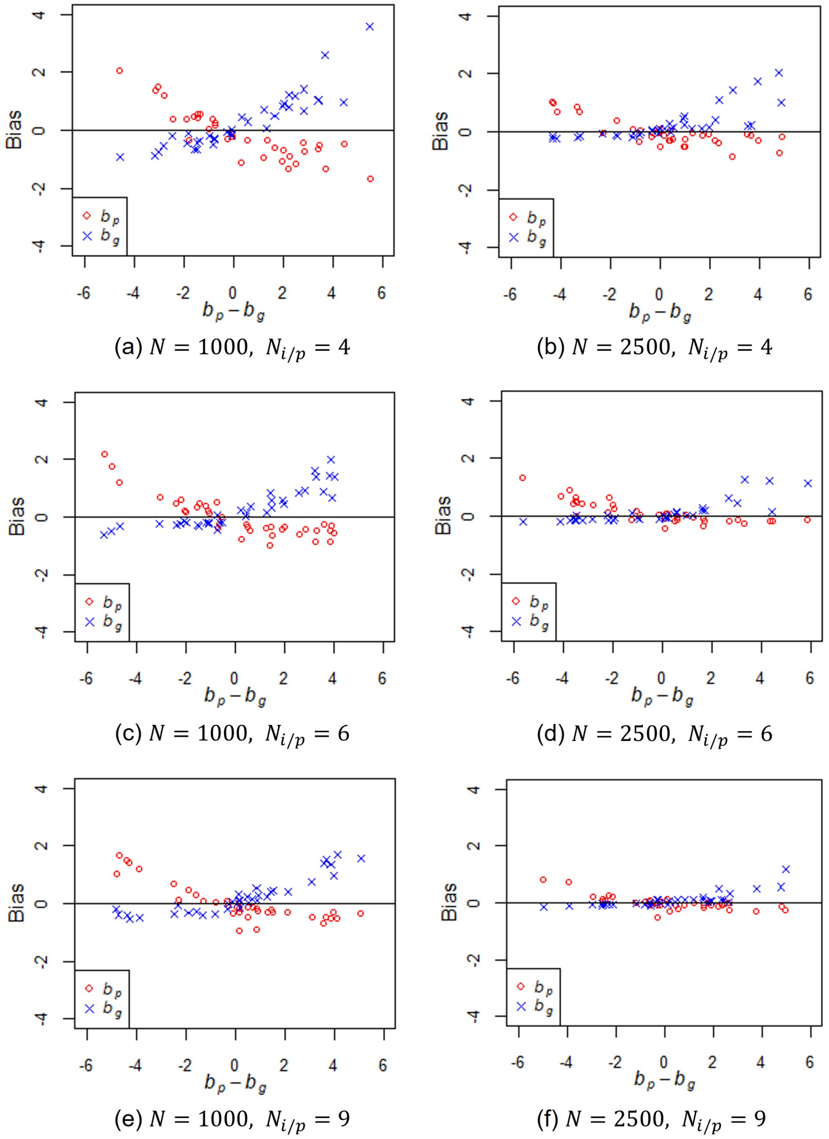
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Appendices

Appendix A Supplementary Figures and Tables for *Noncompensatory MIRT for Passage-Based Tests* Study

Figure A1

Plots of Bias in Difficulty Estimates for Passage-related and Passage-independent Components Under the Passage-Based Noncompensatory Model (PB-NM) Against the Difference in the Two Difficulty Parameters, for Each Generating Condition.



(a) $N = 1000,\ N_{i/p} = 4$

(b) $N = 2500,\ N_{i/p} = 4$

(c) $N = 1000,\ N_{i/p} = 6$

(d) $N = 2500,\ N_{i/p} = 6$

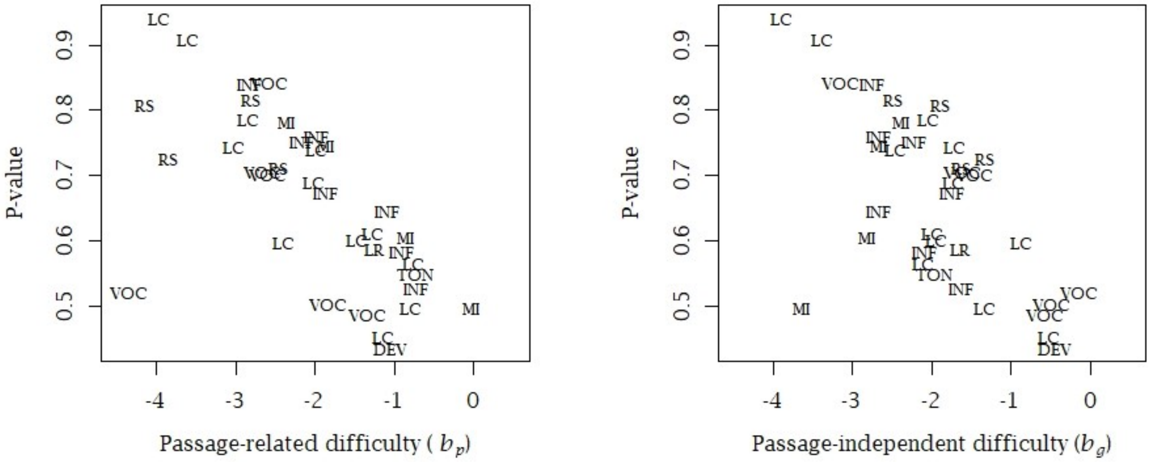
(e) $N = 1000,\ N_{i/p} = 9$

(f) $N = 2500,\ N_{i/p} = 9$

Note. N = The number of respondents; $N_{i/p}$ = The number of items per passage; b_p = Passage-related difficulty; b_g = Passage-independent difficulty.

Figure A2

Plots Showing the Relationship between the Proportion-correct (p-value) and Item Component Difficulty Estimates Under the Passage-Based Noncompensatory Model (PB-NM) with Item Type Labels, for Reading Comprehension Test Data



Note. LC = Literal comprehension; MI = Main idea; INF = Inference; VOC = Vocabulary in context; RS = Rhetorical strategy; DEV = Development; TON = Tone; LR = Logical reasoning.

Table A1

A Model Fit Comparison across Rasch, Two-parameter Logistic (2PL), and Bifactor Models, for Reading Comprehension Test Data.

Model	AIC	BIC	HQ
Rasch	96558.40	96773.90	96636.64
2PL	96238.29	96657.62	96390.53
Bifactor	95902.23	96531.23	96130.59

Note. AIC = Akaike Information Criterion ; BIC = Bayesian Information Criterion; HQ = Hannan-Quinn Criterion.

Appendix B Examples of Items for Each Item Type in the Reading Comprehension Test

- **Literal comprehension (LC)**

  The passage claims that seeing is

  A. something our brains do much more effectively than thinking.

  B. something at which humans excel other animals.

  C. an important prerequisite to thinking.

  D. more important than thinking.

  E. closely related to our power of memory.

- **Main idea (MI)**

  The main point of this passage is that barbed wire fences

  A. hurt the American economy.

  B. ruined the land in the American West.

  C. made a huge impact on the make-up of the American West.

  D. made inventors rich.

  E. caused problems between settlers.

- **Inference (INF)**

  The author seems to imply that humans

  A. are much like animals.

  B. do not spend much time thinking.

  C. have the ability to reason.

  D. cannot reason well.

  E. rely on memory.

- **Vocabulary in context (VOC)**

  The word "analogy" in line 13 most nearly means

  A. experiment.

  B. function.

  C. structure.

  D. comparison.

  E. cause.

- **Rhetorical strategy (RS)**

  The first paragraph of this selection serves what purpose?

A. To introduce Edgar Allan Poe's most famous stories.

B. To offer a definition of "detective."

C. To explain characteristics of the detective story genre.

D. To get readers interested in reading the stories of Edgar Allan Poe.

E. To use expert opinion to set up the claims in the second paragraph.

- **Development (DEV)**

This passage is composed primarily of

A. comparisons.

B. examples.

C. descriptions.

D. generalizations.

E. factual claims.

- **Tone (TON)**

The tone of the passage is

A. respectful.

B. objective.

C. bitter.

D. ironic.

E. sarcastic.

- **Logical reasoning (LR)**

There is no item that can be exposed to public for this item type. A common form for this item type is to ask examinees to identify an unstated assumption the author is making based on a claim that is stated in the passage.

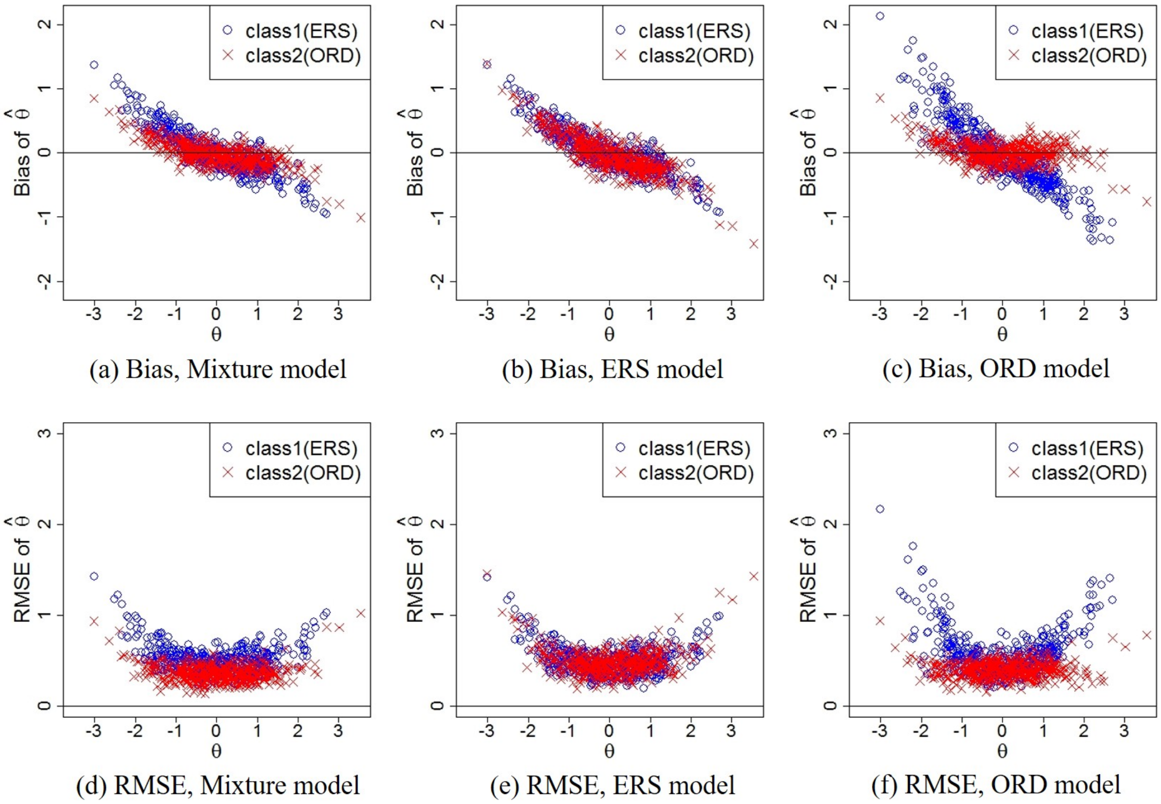
Appendix C Recovery of Respondent Parameters in the Simulation: Bias and Root Mean Squared Error (RMSE) of Respondent Parameter Estimates

The mixture IRTree model also recovers the person parameters θ and η reasonably well across classes. For the simulation reported in the study, Figures C1 and C2 show the bias and RMSE of θ and η estimates, respectively, produced from the mixture and single IRTree models for the condition where the proportion of each class is simulated as equal (i.e., (P_1, P_2) = (0.5, 0.5)). As a result, the true data generated contain both ORD and ERS respondents. Looking first at Figure C1, the results are shown in terms of bias of the θ estimate as well as RMSE, when each of the three different models (mixture, ERS only, ORD only) is fit. Note that for respondents in Class 1 (the ERS class), the bias and RMSE of the estimates derived from the mixture and ERS model are similar whereas the ORD model produces larger errors. This can naturally be explained by the fact that the ORD model does not account for extreme response style. For respondents in Class 2 (the ORD class), by contrast, the mixture model produces similar recovery to that of the ORD model while the ERS model produces slightly larger errors. This can be explained by the fact that the ERS model uses information from only one node of each item to estimate θ, resulting in much less information, whereas the other two models use all three nodes.

As to the η estimates, I focus on the estimates observed under the mixture and ERS models only. Figure C1 shows that the overall patterns of bias and RMSE from the mixture and ERS models appear to be similar while the mixture model produces slightly better results. As can be seen from the plots, for Class 1, the points are scattered slightly more closely to zero for the mixture model than for the ERS model. For Class 2, both the mixture and ERS models show large bias and RMSE of estimates (it is viewed that the true ERS parameters for respondents in Class 2 are still present even though their responses to items are not influenced by the ERS). The errors get larger as the deviation of η from zero increases because the η estimates for these respondents from both models are close to zero, which can be viewed as a reasonable estimate for η based on the fact that their responses are not affected by ERS. The points from the mixture model seem to be scattered almost as a straight line (reflecting that all the estimates are nearly 0) whereas the points from the ERS model are more spread. It can reasonably argued from such results that the mixture model classifies respondents to their true class pretty well, and therefore, the person parameters θ and η can be estimated as accurately as (or perhaps slightly better than) estimated under the correct single IRTree model.

Figure C1

Bias and RMSE of Content Trait (θ) Estimates Obtained From the Mixture IRTree Model [(a), (d)], Single ERS IRTree Model [(b), (e)], and Single ORD IRTree Model [(c), (f)], for (P_1, P_2) = (0.5, 0.5) Condition

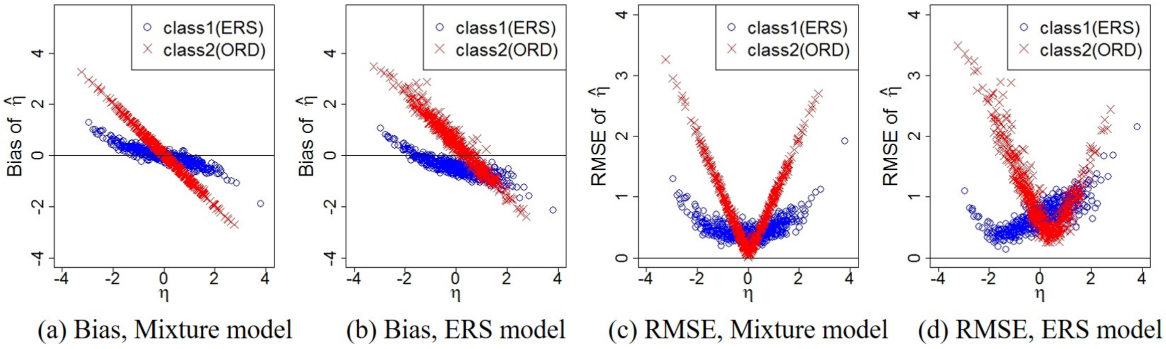


(a) Bias, Mixture model (b) Bias, ERS model (c) Bias, ORD model

(d) RMSE, Mixture model (e) RMSE, ERS model (f) RMSE, ORD model

Figure C2

Bias and RMSE of ERS (η) Estimates Obtained From the Mixture IRTree Model [(a), (c)] and Single ERS IRTree Model [(b), (d)], for (P_1, P_2) = (0.5, 0.5) Condition



(a) Bias, Mixture model (b) Bias, ERS model (c) RMSE, Mixture model (d) RMSE, ERS model

I note that these results in terms of recovery under model misspecification are very likely affected by the psychometric characteristics of items used to generate the data. In particular, the item parameters simulated in this study imply that tests are centered with respect to the rating scale (i.e. equivalent numbers of “agree” and “disagree” responses). The results will likely imply greater bias under misspecification when the items are not centered (e.g., a disproportionate number of either “agree” or “disagree” responses), as is common in practice on self-report survey instruments.