Using Item Response Tree Models for Studying Response Behaviors

by

Minjeong Park

B.A., Ewha Womans University, 2014

M.A., The University of British Columbia, 2017

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

November 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Using Item Response Tree Models for Studying Response Behaviors

submitted by Minjeong Park in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Measurement, Evaluation, and Research Methodology

Examining Committee:

Dr. Amery D. Wu, Associate professor, Faculty of Education, UBC
Supervisor

Dr. Sterett H. Mercer, Professor, Faculty of Education, UBC
Supervisory Committee Member

Dr. Farinaz Havaei, Assistant Professor, School of Nursing, UBC
Supervisory Committee Member

Dr. Gunderson Lee, Professor, Faculty of Education, UBC
University Examiner

Dr. Jeremy Biesanz, Associate professor, Faculty of Arts, UBC
University Examiner

Abstract

In psychological and educational testing, test-takers' response behaviors are a critical issue because they have significant impacts on the measurement of the trait and ability. The item response tree (IRTree) model is recently introduced as a promising tool for studying response behaviors. In this dissertation, I focused on the *explanatory IRTree model* that allows the researchers to include person and item characteristic variables to "explain" response behaviors. Although the explanatory IRTree model provides a useful way to address various queries about response behaviors, it has not gained much attention in the literature. Thus, the goal of this dissertation is to draw researchers' attention to the potential of the explanatory IRTree model. To do so, I first introduced the IRTree model within an explanatory item response modeling framework. Taking this framework, I explicated how the standard IRTree model (a.k.a., descriptive IRTree model) can be easily extended to the explanatory IRTree model.

Following that, I showcased two real-data applications. Study-1 used both the descriptive and explanatory IRTree models to inspect the response styles when answering the Rosenberg's Self-esteem Scale. As the main findings, this study found the presence of two distinct extreme response styles and the acquiescence response style. Study-2 used the explanatory IRTree model to investigate the effects of person and item characteristics on nonresponse behaviors (not-reached and omitted) when taking the reading test of the Progress in International Reading Literacy Study. The findings showed that item nonresponse behaviors occurred differently depending on gender, test language, item location, and item format.

There are three unique contributions of this dissertation. First, it expanded the utility of the IRTree model to be a tool for "explaining" response behaviors. Second, it provided an in-depth understanding of response styles in Likert-type psychological rating scales and

nonresponse behaviors in educational testing. Finally, the method and findings of the two studies offered practical implications on the test/scale development and validation.

Lay Summary

In psychological and educational testing, individuals' response behaviors can have significant impacts on the meaning of the measured scores. The item response tree (IRTree) model is recently introduced as a promising tool for studying individuals' response behaviors. This dissertation expanded the standard IRTree model to become a tool for explaining response behaviors by introducing the *explanatory* IRTree model. Moreover, I showcased its applications in the context of studying two common response behaviors: response styles and item nonresponses. Study-1 investigated response styles (i.e., individuals' tendencies to prefer or avoid particular response categories) when answering the psychological rating scale called Rosenberg's Self-esteem Scale. Study-2 examined nonresponse behaviors (i.e., omitting the questions, not completing the test) in taking the reading test called Progress in International Reading Literacy Study. Through these applications, this dissertation showed the benefits of the explanatory IRTree model and discussed the implications of the findings from the two studies.

Preface

This dissertation is an original intellectual product of the author, Minjeong Park.

A version of Chapter 3 has been published in a peer-reviewed journal. The citation in APA style is given below:

> 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.

I was the lead investigator of this research article, responsible for formulating the research questions, data analyses, and the writing of most of the article. Dr. Wu, my dissertation supervisor, provided guidance and feedback throughout the entire project from generating the initial idea to the final publication.

Chapter 4 was written as a stand-alone manuscript. Dr. Wu and I played similar roles to those in the article of Chapter 3, except that this manuscript has not yet been published, nor was it peer-reviewed by any journals.

I did not collect any data for this dissertation. Study-1 and Study-2 were conducted based on publicly available data for secondary use. No individual identification was revealed in these two datasets. Thus, no ethics approvals were needed. The first dataset was retrieved from the 2005 Longitudinal Study of Generation in California (Silverstein & Bengtson, 2008) and the second dataset was retrieved from the Progress in International Reading Literacy Study 2016 (Mullis, Martin, Foy, & Hooper, 2017).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

-2 LL: -2 times log likelihood

AIC: Akaike information criterion

BIC: Bayesian information criterion

GLNMM: Generalized linear and nonlinear mixed model

IRT: Item response theory

IRTree: Item response tree

SEM: Structural equation modeling

VIF: Variance inflation factor

Acknowledgments

Throughout my dissertation journey, I have received a great deal of support.

I would like to thank my supervisor Dr. Amery D. Wu for her support, encouragement, and guidance during my studies. The courses, meetings, and conversations with you inspired me to continue my work in this field and complete this dissertation. I also wish to thank my committee members, Dr. Sterett H. Mercer and Dr. Farinaz Havaei, for their support and critical feedback.

I would like to give my gratitude to the faculty, fellow students, and coworkers I have met during my studies, who have given me great learning opportunities and memories at UBC. All these experiences meant a great deal to me.

Special thanks are owed to my family for all the support and encouragement you have shown me throughout my studies.

And my biggest thanks to Stephen for your enduring support and encouragement. With you, my journey was a lot smoother and joyful. For my little furry friend Nana, thank you for always cheering me up and giving me happy distractions to rest and comfort my mind.

Dedication

In memory of my grandmother

Chapter 1: Introduction

Psychological and educational tests have been widely used to measure test-takers' underlying traits and abilities. In many disciplines, psychological tests are developed to measure personal traits such as personality, attitude, interest, and psychological status. Well-developed psychological tests have contributed to decision-making in various settings including diagnosis, intervention, judicial and government decisions, and personal awareness (AERA et al., 2014). By contrast, educational tests are commonly adopted to assess test-takers' learning, knowledge, and skills. This type of test ranges from classroom assessments to large-scale standardized tests. Examples include the Scholastic Aptitude Test, American College Testing, Progress in International Reading Literacy Study, and National Assessment of Educational Progress. The results of these educational tests are widely used to make judgments about students' learning progress and provide valuable insights into teaching, performance development, and educational policy (AERA et al., 2014; Cresswell et al., 2015).

In psychological and educational testing, test-takers' response behaviors are a critical issue because they have a significant implication on the measurement of trait and ability. In this dissertation, I will use the term response behavior to refer to the way in which test-takers respond to the test. Specifically, unintended/undesirable response behaviors have been intensively investigated because they can cause serious biases in the measurement. Such examples include omitting the items (e.g., Di Chiacchio et al., 2016; Köhler et al., 2015), misunderstanding of the questions/statement (e.g., Barrett, 2004; Finlay & Lyons, 2001; Lavrakas, 2008b), preferring/avoiding particular response categories (e.g., Böckenholt, 2017; Khorramdel & von Davier, 2014), responding carelessly (e.g., Meade & Craig, 2012; Rios et al., 2017), providing

dishonest responses (e.g., Austin, 1992; Foulds & Warehime, 1971), and guessing answers (e.g., Parker & Ryan, 1993; Pokropek, 2016). In test development and validation, identifying these unintended/undesirable response behaviors and their adverse effects on the measurement provides valuable insight.

1.1 The focus of this research

In psychometrics, response behaviors have been commonly investigated by applying traditional item response models such as the Rasch model (Rasch, 1960), the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), and the nominal response model (Bock, 1972). Recently, other types of item response models have been used to investigate response behaviors. One such model is the item response tree (IRTree) model. The IRTree model (also known as multi-process item response theory model or multinomial processing tree model) is a tree-based item response model introduced by Böckenholt (2012) and De Boeck and Partchev (2012). In essence, this approach enables researchers to postulate a decision process in a tree structure and analyze the data accordingly. For example, in a test including four response categories of 'Strongly agree' to 'Strongly disagree' with no neutral/middle point, researchers can hypothesize the following decision process: (1) test-takers decide whether they agree (i.e., choose either of the agree categories) or disagree (i.e., choose either of the disagree categories) with the statement and (2) test-takers then determine how strongly they agree or disagree (i.e., choose extreme categories or mild categories). In doing so, the IRTree model allows researchers to examine the underlying trait/ability and item parameters associated with different decision steps. A detailed overview of the IRTree model will be provided in Chapter 2.

Since the introduction of the IRTree model in 2012, there had only been periodic yet infrequent publications on it. To my best knowledge, I had exhausted all the records I could find at the time of conducting the literature review for the dissertation, using the keyword 'IRTree' and its synonyms – 'item response tree model', 'multi-process item response theory model', and 'multinomial processing tree.' That said, I am aware that the IRTree model is gaining much more attention today, and there may be new publications that have not been included in this dissertation.

In the literature, the IRTree model has been considered a promising tool for studying various response behaviors. In particular, it has gained popularity in the study of response styles (i.e., test-takers' tendencies to prefer/avoid a particular type of response categories such as extreme categories). By explicitly modeling the process of choosing response categories ('Strongly agree' to 'Strongly disagree') in the IRTree model, previous studies have examined the tendencies to use extreme and mid-point response styles (e.g., Böckenholt, 2012; Khorramdel & Davier, 2014; Plieninger & Meiser, 2014; Zettler et al., 2016).

Meanwhile, other studies showed that the IRTree model is useful for investigating response behaviors that occur sequentially. For example, Jeon et al. (2017) modeled test-takers' answer change behaviors in a math test (e.g., changing from the incorrect answer to correct answer; from the correct answer to incorrect answer). By modeling it as a process via an IRTree model, they tested the underlying abilities associated with different changing patterns. Moreover, some research analyzed test-takers' fast and slow response behaviors based on the two-step process in an IRTree model (DiTrapani et al., 2016; Partchev & De Boeck, 2012). They first specified whether test-takers provided fast or slow responses and then whether they provided a correct or incorrect response for fast and slow responses, respectively. In doing so, they

examined the underlying abilities differentiated by response time. Likewise, several studies examined test-takers' nonresponding behaviors (e.g., omitting) by postulating the sequential process of producing missing and valid responses in educational tests (Debeer et al., 2017; Okumura, 2014). More recently, the IRTree model was also applied in other various contexts such as raters' scoring processes (Myers et al., 2020), eye-tracking data (Cho et al., 2020), and response processes (LaHuis et al., 2019).

The IRTree model can further provide great opportunities to study response behaviors when incorporating person and item characteristics to explain response behaviors. In this dissertation, I will call an IRTree model with person and item characteristics an *explanatory IRTree model*. In the explanatory IRTree model, the researchers can explain the systematic effects of test-taker and item characteristics on the decision process, over and beyond the trait/ability that the test intends to measure. This can extend the use of the IRTree model to address substantive research questions about response behaviors. For example, it helps researchers examine whether and how response behaviors occur differently depending on person characteristics such as gender, ethnicity, and education level. Similarly, researchers can test whether test-takers respond to the items differently depending on item features such as item format. Thereby, it can greatly enhance our understanding of response behaviors.

To date, the literature regarding the explanatory IRTree model is still scant. As a pioneer work, Okumura (2014) applied the explanatory IRTree model and examined how omitting response behavior occurred differently depending on factors such as item format, sex, and enjoyment of reading. More recently, two studies used the explanatory IRTree model to examine how test-takers' response patterns were affected by item features, such as item keying (reversely scored or not) and item wording (positively or negatively wording) (Böckenholt, 2019; Wu &

Jin, 2020). Meanwhile, Jeon and De Boeck (2016) and Debeer et al. (2017) pointed out that a standard IRTree model can be generally extended to an explanatory model. However, despite these previous works, the explanatory IRTree model has not gained much attention and remained, by and large, unexplained to the applied researchers.

1.2 Purpose of this research

This research aims to draw researchers' attention to the explanatory IRTree model and its potential in the study of response behaviors. This approach can provide unique opportunities to address substantive research questions by incorporating person and item characteristics in the IRTree model. However, such a possibility was hardly discussed in the literature. In this dissertation, I will elaborate on the explanatory IRTree model and shed light on its usefulness. To do so, this research includes the following specific objectives.

First, I will introduce the IRTree model within a large psychometric framework, *explanatory item response modeling*, proposed by De Boeck and Wilson (2004). This framework enables the researchers to go beyond the standard item response models (called descriptive models) and formulate the explanatory models in a straightforward way. Because of this benefit, the explanatory item response modeling framework has been adopted in various applications (e.g., Briggs, 2008; Hartig et al., 2012; Min et al., 2018; Randall et al., 2011; Stanke & Bulut, 2019). By the same token, the IRTree model can take benefit from this framework. However, to my knowledge, this framework has not been applied to the IRTree model. In this dissertation, I will introduce the IRTree model in this framework and explicate how the standard descriptive IRTree model can be easily extended to the explanatory IRTree model.

Second, this dissertation will show two novel applications of the IRTree model with empirical data. In Study-1, I will showcase the application for studying response styles. Response styles are individuals' tendencies to prefer or avoid particular response categories (e.g., extreme categories), which are commonly observed in psychological measurement (Cronbach, 1946; Jackson & Messick, 1958; Paulhus, 1991). In this study, I will use both the descriptive and explanatory IRTree models for investigating extreme, acquiescence, and disacquiescence response styles in the Rosenberg's Self-esteem Scale. In Study-2, I will focus on item nonresponse behaviors. In large-scale educational tests, it is not uncommon that test-takers omit the items or fail to complete the test. Such behaviors are generally referred to as item nonresponse (De Leeuw et al., 2003; Groves, 1989; Huisman, 1999; Köhler et al., 2017). By applying the explanatory IRTree model, I will examine the effects of person and item characteristics on two types of item nonresponse (not-reached and omitted responses) in the Progress in International Reading Literacy Study.

1.3 Organization of this research

There are five chapters in this dissertation. Chapter 1, the current chapter, introduces the background, research purpose, and an overview of the dissertation. Chapter 2 will explicate the IRTree model under the explanatory item response modeling framework. In this chapter, I will first introduce the standard IRTree model with an illustrative example. Then, I will review the explanatory item response modeling framework. Based on that, I will show how this framework subsumes the IRTree model and extend it to the explanatory IRTree model.

In Chapter 3 and Chapter 4, I will introduce the two studies applying the IRTree model. Chapter 3 will present Study-1 titled "Item response tree models to investigate acquiescence and

extreme response styles in Likert-type rating scales." This chapter is a journal article published in the peer-reviewed journal *Educational and Psychological Measurement*. In this chapter, I will first introduce response styles and provide a review of the related statistical approaches and the response styles of interest. Following that, I will discuss the specification of the descriptive and explanatory IRTree models for examining the extreme, acquiescence, and disacquiescence response styles. These models will then be demonstrated with the Rosenberg's Self-esteem Scale.

Chapter 4 will present Study-2 which is titled "Examining person and item characteristics associated with not-reached and omitted responses using item response tree models." It is written in a manuscript format for publication. This chapter will begin by introducing item nonresponse behaviors in large-scale educational tests. I will then provide a literature review on person and item characteristics related to item nonresponse and different statistical approaches. Next, I will explain the specification of the explanatory IRTree model to investigate the two types of item nonresponse (not-reached and omitted) and the factors associated with them. This application will be demonstrated with the data from the Progress in International Reading Literacy Study 2016.

Lastly, Chapter 5 will provide a discussion. This chapter will first summarize the entire work presented in this dissertation. I will then highlight the contributions and novelties. Following that, I will talk about several issues that may arise in fitting the IRTree model, in particular, when using the *lme4* R package. Finally, the limitations of the current work and the recommendations for future research will be discussed.

Note that Chapters 3 and 4 are written in a manuscript format. Although Chapter 3 has already been published, I have made minor modifications for better flow with the rest of this

dissertation. This includes removing redundant content and adding extra comments. Likewise, Chapter 4 is written to submit for publication and I removed the redundant content for fluency.

Chapter 2: IRTree Models within a Large Psychometric Framework

In this chapter, I will introduce the IRTree model under a large psychometric framework, called explanatory item response modeling, introduced by De Boeck and Wilson (2004). To facilitate readers' understanding, I will first provide an overview of the standard IRTree model with an illustrative example. Thereafter, I will introduce the explanatory item response modeling framework and then explain how the standard IRTree model can be extended to the explanatory IRTree model within this framework.

2.1 Overview of the standard IRTree model

A variety of item response models (also referred to as item response theory models) have been introduced to analyze categorical item responses. Most well-known examples include the Rasch model (Rasch, 1960), the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), and the nominal response model (Bock, 1972). The IRTree model, as a new type of item response model, was introduced by Böckenholt (2012) and De Boeck and Partchev (2012). The IRTree model differs from the traditional item response models in that it enables the users to postulate a decision process in item responding. This feature is particularly beneficial to study various response behaviors involved in the process. In this section, I will give an introduction to the standard IRTree model. In the following, I will first describe an illustrative example and, based on that, explain the key components of the IRTree model including the tree structure, mapping matrix, and model specification.

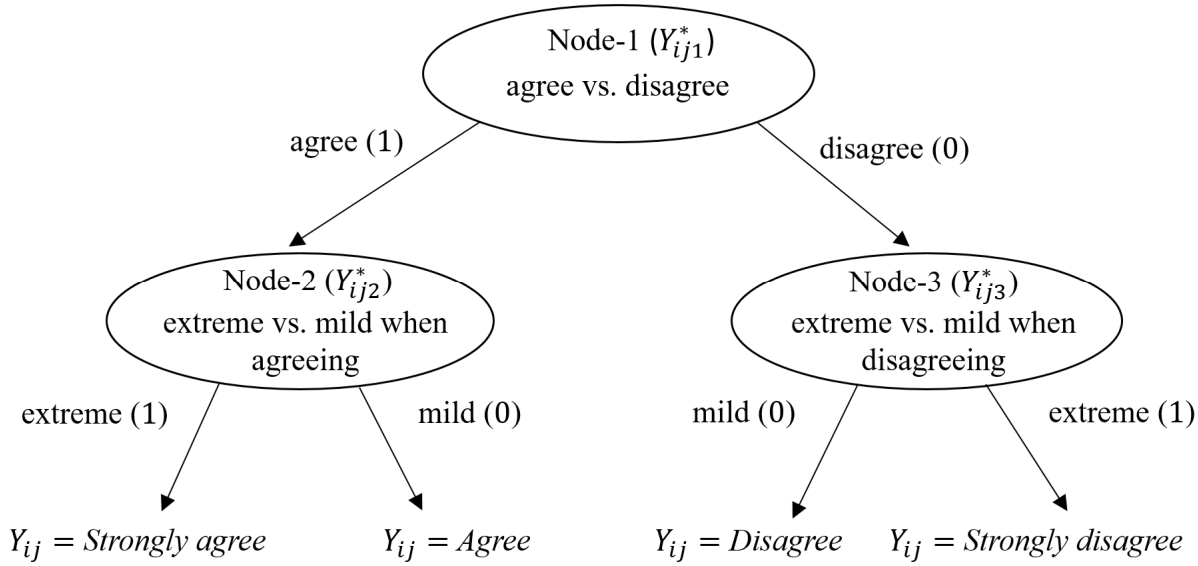
In psychological tests, test-takers are often asked to choose from a set of response categories indicating different levels of agreement (e.g., Strongly agree to Strongly disagree).

This response format is often referred to as a Likert-type rating scale. Suppose that a four-point Likert-type rating scale comprises of Strongly agree (SA), Agree (A), Disagree (D), and Strongly disagree (SD) response categories. Within this scale, researchers can postulate that individuals' item responses are derived from a decision process. As an example, the following decision process can be specified: (1) the respondents first decide whether they agree or disagree with an item statement, and (2) they then determine how strongly they agree or disagree with the statement.

2.1.1 Tree structure

In the standard IRTree model, this decision process can be depicted by a series of *nodes* and *branches* in a tree structure as shown in Figure 2.1. Each node represents a decision query, and the branches represent the decisions made at each node. Note that the nodes are sometimes called pseudo items or sub-items in the literature (e.g., Böckenholt, 2017; De Boeck & Partchev, 2012). In Figure 2.1, Node-1 represents whether the agree categories (SA and A) or disagree categories (SD and D) are chosen (agree = 1, leading to the left branch; disagree = 0, leading to the right branch). If the decision at the first node is 1 (i.e., A or SA is chosen), Node-2 represents whether the extreme category (SA) or mild category (A) is chosen (extreme = 1 leading to the left branch; mild = 0 leading to the right branch). If the decision at the first node is 0 (i.e., D or SD is chosen), Node-3 represents whether the extreme category (SD) or mild category (D) is chosen (extreme = 1 leading to the right branch; mild = 0 leading to the left branch).

Figure 2.1 An example tree structure for a four-point Likert-type rating scale.



2.1.2 Mapping matrix

Following the hypothesized tree structure, the IRTree model re-represents each observed response (i.e., SA, A, D, SD) by a set of outcomes for Node-1, Node-2, and Node-3 as shown in Table 2.1. This table is referred to as a *mapping matrix* (De Boeck & Partchev, 2012). For example, the observed response 'Agree' is represented by (1, 0, NA) for the three nodes. Note that the irrelevant decision query is recoded as not applicable (NA). As an example, for the observed responses 'Agree' and 'Strongly agree', Node-3 (choosing between the disagree categories of SD and D) is irrelevant to the process. Therefore, Node-3 is coded as NA for these responses. For the observed responses 'Disagree' and 'Strongly disagree', the situation is reversed. Node-2 (choosing between the agree categories of SA and A) is irrelevant and therefore coded as NA.

Table 2.1 Mapping matrix for the decision process based on Figure 2.1

Responses (Y_{ij})	Node-1 (Y^*_{ij1})	Node-2 (Y^*_{ij2})	Node-3 (Y^*_{ij3})
Strongly agree	1	1	NA
Agree	1	0	NA
Disagree	0	NA	0
Strongly disagree	0	NA	1

Note. NA denotes not applicable.

By applying the mapping matrix, we can transform the original item response into a data matrix shown in Table 2.2. This table presents some examples of the data values transformed based on the mapping matrix in Table 2.1. For example, in the first row of Table 2.2, person 1's observed response to Item-1 is 'Agree', so this response is transformed to node outcomes (1, 0, NA), shown under the columns for the three nodes.

Table 2.2 An example data matrix according to the mapping matrix

Person	Item (original responses)	Node-1 (Y^*_{ij1})	Node-2 (Y^*_{ij2})	Node-3 (Y^*_{ij3})
1	Item-1 (A)	1	0	NA
1	Item-2 (SD)	0	NA	1
…	…	…	…	…
100	Item-1 (D)	0	NA	0
100	Item-2 (SA)	1	1	NA
…	…	…	…	…

Note. A = agree; SA = strongly agree; D = disagree; SD = strongly disagree; NA = not applicable.

Finally, we should reshape the data matrix into a long format in order to fit the IRTree model, as presented in Table 2.3. The "Node" column indicates the three nodes, and the corresponding outcomes are presented in the "Node outcome" column. For example, the first three rows of node outcomes now represent person 1's 'Agree' response to Item-1.

Table 2.3 Long format of an example data matrix for fitting the IRTree model

Person	Item (original responses)	Node	Node outcome
1	Item-1 (A)	Node-1	1
1	Item-1 (A)	Node-2	0
1	Item-1 (A)	Node-3	NA
1	Item-2 (SD)	Node-1	0
1	Item-2 (SD)	Node-2	NA
1	Item-2 (SD)	Node-3	1
	…		
100	Item-1 (D)	Node-1	0
100	Item-1 (D)	Node-2	NA
100	Item-1 (D)	Node-3	0
100	Item-2 (SA)	Node-1	1
100	Item-2 (SA)	Node-2	1
100	Item-2 (SA)	Node-3	NA
	…		

Note. A = agree; SA = strongly agree; D = disagree; SD = strongly disagree; NA = not applicable.

2.1.3 Model specification

Based on the long format data, we can fit the IRTree model to examine the node-specific parameters as discussed by De Boeck and Partchev (2012) and Jeon and De Boeck (2016). That is, the probability of the outcome for each node Y_{ijn}^{*} can be specified as,

$$\pi\left(Y_{ijn}^{*} = y_{ijn}^{*}\right) = \frac{\exp(\theta_{in}+\beta_{jn})^{y_{ijn}^{*}}}{[(1+\exp(\theta_{in}+\beta_{jn}))]} = g^{-1}\left(\theta_{in} + \beta_{jn}\right), \quad (2.1)$$

where y_{ijn}^{*} is the outcome of node *n* for person *i* on item *j*. The θ_{in} denotes person *i*'s latent trait that is involved in the decision-making at the *n*th node. The β_{jn} indicates the item parameter of item *j* for the *n*th node. The g^{-1} denotes the inverse of the link function (typically a logit or probit link). In this illustrative example, each node has binary outcomes coded as either 0 or 1. Therefore, the probabilities of the binary outcomes at each node are given as,

$$\pi\left(Y_{ijn}^{*} = 0\right) = \frac{1}{[1+\exp(\theta_{in}+\beta_{jn})]} \quad (2.2)$$

$$\pi\left(Y_{ijn}^{*} = 1\right) = \frac{\exp(\theta_{in}+\beta_{jn})}{[1+\exp(\theta_{in}+\beta_{jn})]} \quad (2.3)$$

Then, we can compute the probabilities of the originally observed responses (SA, A, D, and SD) based on the joint probability of three node outcomes (Böckenholt, 2017; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). For example, Table 2.4 presents the probabilities of the four observed responses.

Table 2.4 Probabilities of the observed responses

Responses (Y_{ij})	Probability
Strongly agree	$\pi(Y_{ij} = \text{Strongly agree}) = \pi(Y^*_{ij1} = 1) * \pi(Y^*_{ij2} = 1)$
Agree	$\pi(Y_{ij} = \text{Agree}) = \pi(Y^*_{ij1} = 1) * \pi(Y^*_{ij2} = 0)$
Disagree	$\pi(Y_{ij} = \text{Disagree}) = \pi(Y^*_{ij1} = 0) * \pi(Y^*_{ij3} = 0)$
Strongly disagree	$\pi(Y_{ij} = \text{Strongly disagree}) = \pi(Y^*_{ij1} = 0) * \pi(Y^*_{ij3} = 1)$

So far, I have summarized the standard IRTree model. In the rest of this chapter, I will explicate how we can extend the standard IRTree model to the explanatory IRTree model that includes person and item properties to explain individuals' differences in the decision process. Specifically, I will borrow from the explanatory item response modeling framework articulated by De Boeck and Wilson (2004). In the following section, I will first review the explanatory item response modeling framework and then introduce the explanatory IRTree model within this framework.

2.2 Explanatory item response modeling framework

Traditionally, many item response models focused on analyzing item responses as a function of latent variables (usually denoted by θs) and item parameters such as item difficulty and item discrimination. These traditional item response models do not incorporate person and item characteristics. The explanatory item response modeling framework, proposed by De Boeck and Wilson (2004), extends the traditional item response models to the explanatory models including person and item characteristic variables to examine their effects on the item responses. In the following, I describe two major aspects of this framework.

2.2.1 Generalized linear and nonlinear mixed models

Generalized linear and nonlinear mixed model (GLNMM) is a broad classification of statistical methods that model the categorical data as a function of the predictors (Baayen et al., 2008; De Boeck & Wilson, 2004; Rijmen et al., 2003). This framework takes the view that item response models are a form of GLNMM. Item response models are "generalized" because the categorical item response data are modeled as a function of the predictors via a link function (typically logit or probit). The relationship between the predictors and the categorical response data can be specified as "linear" or "nonlinear". For example, the one-parameter item response models, fixing the item discrimination parameters to be equal across all items, are linear models. In contrast, the two- and three-parameter models that allow item-specific discrimination parameters are nonlinear models. Meanwhile, item response models can be seen as a "mixed" model because they include both fixed and random effects. For example, in the typical item response models, the latent trait variable is the random effects of persons, and the item parameters are fixed effects of items.

By treating item response models as GLNMMs, this framework showed that a variety of item response models can be formulated as regression-like models.

2.2.2 Descriptive and explanatory models

With the basis described above, this framework classifies two types of item response models: *descriptive* and *explanatory models*. The key distinction between these two types of models lies in the attribute of the predictors included in the model – indicator vs. property. A model is descriptive when it includes only the indicator variables (e.g., person ID, item ID) as predictors. For example, a common form of descriptive model treats the person indicator variable

as having a random effect (i.e., leading to a latent variable) and the item indicator variable as having a fixed effect (e.g., leading to item difficulty, item discrimination). In this respect, the widely used traditional one- to three-parameter binary IRT models are descriptive models, and so are other popular polytomous IRT models such as the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), and the nominal response model (Bock, 1972). In contrast, an explanatory item response model includes at least one property variable as a predictor, e.g., person properties of gender and ethnicity or item properties of item format, item keying direction, and item wording. Therefore, the key purpose of the explanatory model is to explain the effects of the person/item properties on the responses. The linear logistic test model by Fischer (1973) and its extensions are an example of such models.

In a nutshell, the explanatory item response modeling framework treats all item response models as GLNMM that either includes property variables (explanatory) or indicator variables (descriptive) as predictors for the categorical response outcomes. In doing so, this framework analyzes item response data as a regression-like model that includes a combination of different types of predictors to have fixed or random effects. This perspective is useful to straightforwardly extend a descriptive model to be an explanatory model so as to "explain" the effects of item/person properties on item responses. This flexibility permits the researchers to specify the most suitable models for their research hypotheses and the data at hand. With these advantages, explanatory item response models have been slowly becoming known and used in the literature (e.g., Briggs, 2008; Hartig et al., 2012; Min et al., 2018; Randall et al., 2011; Stanke & Bulut, 2019).

Since the standard IRTree model is a type of descriptive item response model, we can easily extend it to be an explanatory model by the same token. In the next section, I will explain

how this framework subsumes the IRTree model and explicate the explanatory IRTree model through GLNMM.

2.3 Introducing the explanatory IRTree model

The IRTree model can be subsumed under the explanatory item response modeling framework. The IRTree model as an item response model fits well into the broader statistical theory of GLNMM. That is, the standard IRTree model shown in equation (2.1) analyzes the categorical node outcomes Y_{ijn}^{*} as a function of person and item indicator variables as predictors. Specifically, it treats the person indicators as having random effects (representing the node-specific latent trait θ_{in}), and the item indicators as having fixed effects (representing the node-specific item parameters β_{jn}). Through a link function, the categorical node outcomes are modeled as a linear combination of these indicator predictors. Thus, the standard IRTree model can be perceived as a regression-like model and also classified as a descriptive model because it only includes person and item indicators as predictors.

The standard IRTree model turns into an explanatory model by including any person or item property variables as predictors. This allows researchers to test hypothesized effects of item/person property on the decision process. As a GLNMM, it is straightforward to include property variables to "explain" node outcomes according to the researcher's hypothesis. More formally, we can express an explanatory IRTree model as,

$$\pi\left(Y_{ijn}^{*} = y_{ijn}^{*}\right) = g^{-1}\left(\theta_{in} + \beta_{jn} + \sum_{g}^{G} \beta_{gn} Z_{ijg}\right). \quad (2.4)$$

Extending equation (2.1), this equation adds a linear combination $\sum_{g}^{G} \beta_{gn} Z_{ijg}$. The newly added combination is the weighted sum of a set of property predictors, where Z_{ijg} is the person i's and

item *j*'s value on the *g*th property predictor (g = 1, 2, …, G); the weight β_{gn} is the slope regression parameter of the *g*th property predictor for the *n*th node. This predictor can be specified as fixed or random effects. Therefore, the probabilities of the binary outcomes at each node can be now calculated as follows.

$$\pi\left(Y_{ijn}^{*}=0\right)=\frac{1}{\left[1+\exp\left(\theta_{in}+\beta_{jn}+\sum_{g}^{G}\beta_{gn}Z_{ijg}\right)\right]} \tag{2.5}$$

$$\pi\left(Y_{ijn}^{*}=1\right)=\frac{\exp\left(\theta_{in}+\beta_{jn}+\sum_{g}^{G}\beta_{gn}Z_{ijg}\right)}{\left[1+\exp\left(\theta_{in}+\beta_{jn}+\sum_{g}^{G}\beta_{gn}Z_{ijg}\right)\right]} \tag{2.6}$$

For the explanatory IRTree model, item/person properties should be additionally included in data preparation. Table 2.5 shows an example data matrix including sex and item type as property variables. In this table, for each person and item, corresponding values of sex and item type are indicated. For example, the rows from person 1 are coded as 'Male' in the sex column, and the rows from Item-1 are coded as 'positively keyed' in the item type column. This data matrix is then reshaped into a long format similar to Table 2.3 to fit the explanatory IRTree model.

Table 2.5 An example data matrix applying the mapping matrix with person and item characteristics

Person	Item (original responses)	Sex	Item type	Node-1 (Y^*_{ij1})	Node-2 (Y^*_{ij2})	Node-3 (Y^*_{ij3})
1	Item-1 (A)	Male	PK	1	0	NA
1	Item-2 (SD)	Male	NK	0	NA	1
…	…	…	…	…	…	…
100	Item-1 (D)	Female	PK	0	NA	0
100	Item-2 (SA)	Female	NK	1	1	NA
…	…	…	…	…	…	…

Note. A = agree; SA = strongly agree; D = disagree; SD = strongly disagree; NA = not applicable; PK = positively keyed item; NK = negatively keyed item.

2.4 Summary

This chapter introduced the IRTree model and elaborated on how the standard IRTree model can be extended to an explanatory IRTree model, under the overarching view of explanatory item response modeling. In the following two chapters, I will present two applications of the IRTree model for studying response behaviors with real data. Specifically, Chapter 3 will showcase the application of both descriptive and explanatory IRTree models for investigating extreme, acquiescence, and disacquiescence response styles in the four-point Likert-type Rosenberg's Self-esteem Scale. Chapter 4 will introduce the application of the explanatory IRTree model to examine the effects of test-takers' characteristics and item features

on two types of item nonresponse (not-reached and omitted) in the Progress in International Reading Literacy Study.

Chapter 3: Item Response Tree Models to Investigate Acquiescence and Extreme Response Styles in Likert-type Rating Scales

3.1 Introduction

A Likert-type rating scale is widely used in many disciplines to measure individual differences in attributes, attitudes, or traits. In this type of scale, the response categories are written to represent different levels of endorsement (e.g., 'Strongly agree' to 'Strongly disagree'). Despite the wide uses of the Likert-type rating scales, this response format has been a concern because the respondents may tend to prefer or avoid particular categories, regardless of the levels of the trait being measured. This phenomenon has been referred to as response style, response set, or response bias in the literature (Cronbach, 1946; Jackson & Messick, 1958; Paulhus, 1991). In this study, we used the term "response style" to express this phenomenon.

The adverse effects of response styles have been widely discussed elsewhere (e.g., Kam & Fan, 2017; Moors, 2012; Weijters et al., 2010). The presence of response styles can cause biases in the measurement of the true trait and also affect the meaning of scores. For example, if the respondents prefer the extreme categories (e.g., 'Strongly agree' and 'Strongly disagree'), their responses can overrepresent or underrepresent the true level of the trait, and therefore the scores are possibly biased. In more extreme cases, the scale scores may be seriously biased by response styles, and they cannot be interpreted as representing the trait of interest. Furthermore, response styles may distort the associations among variables measured by the scales because the biased scale scores can deflate or inflate the correlations among the variables. Due to these undesirable effects, there have been many reports investigating the presence of response styles

(e.g., Hurley, 1998; Meisenberg & Williams, 2008; Moors, 2008, 2012; Schneider, 2016; Weijters et al., 2010). Among them, the most commonly reported are the acquiescence response style, disacquiescence response style, extreme response style, and mid-point response style. A comprehensive summary of these response styles can be found in Baumgartner and Steenkamp (2001) and Van Vaerenbergh and Thomas (2013).

Generally, response styles have been examined by two different approaches depending on how the response styles are captured. The first approach incorporates items that are *external* to the substantive trait being measured in order to track the response style of interest (Greenleaf, 1992b; Weijters et al., 2010). The other approach utilizes only the *internal* items of a scale that are originally designed to measure the substantive trait (e.g., Bolt & Johnson, 2009). This approach does not require extra measures or items. Individuals' response patterns to the internal items are inspected to capture the response styles. Both approaches are equally common and sometimes used concurrently (e.g., Wetzel & Carstensen, 2017).

In addition to the ways of capturing response styles, different statistical techniques were applied to investigate response styles. The simplest is to look at descriptive statistics such as frequency counts, mean, and standard deviation of the item scores (Bachman & O'Malley, 1984; Reynolds & Smith, 2010). Although relatively straightforward, descriptive statistics are not very illuminating because this approach cannot tease apart the response styles from the trait being measured. This makes it hard for researchers to inspect whether the responses reflect the response styles, true traits, or both. Due to this limitation, this approach was only recommended when researchers can include external items to detect response styles (Greenleaf, 1992a).

Other more advanced techniques were proposed within two major modeling frameworks: structural equation modeling (SEM) and item response theory (IRT) model. With SEM, response

styles were often modeled as continuous latent variables using confirmatory factor analysis (e.g., Billiet & McClendon, 2000; Welkenhuysen-Gybels et al., 2003). At times, response styles were modeled as categorical latent variables, and latent class analysis was applied to identify subgroups of individuals who display different preferences/avoidances when selecting the response categories (e.g., Moors, 2003, 2010; Van Rosmalen et al., 2010). As for the IRT model approach, some studies proposed a multidimensional nominal response model to examine and control for the extreme response style (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010). For others, polytomous IRT models such as the partial credit model were extended to mixture models to identify latent groups of distinct response styles (e.g., Austin et al., 2006).

Recently, a tree-structure based item response model, item response tree (IRTree) model, gained popularity in the study of response styles (e.g., Böckenholt, 2017; Böckenholt & Meiser, 2017; Khorramdel & Davier, 2014; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013; Zettler et al., 2016). By applying the IRTree model, researchers can explicitly specify the process of choosing the response categories as a series of multiple decision queries. In doing so, it provides great flexibility in investigating various tendencies to prefer/avoid particular response categories based on latent variables and item parameters associated with different decision steps. This feature also enables researchers to disentangle response styles from the substantive trait, based only on the internal items. Due to these benefits, previous studies have applied the IRTree model for examining response styles. However, most of these pioneer studies, if not all, focused on only the extreme and/or mid-point response styles in a five-point rating scale. There was little work on how the IRTree model can be extended to investigate other response styles.

This paper aims to show the application of the IRTree model to investigate a variety of hypotheses about response styles. Specifically, this study will examine extreme, acquiescence, and disacquiescence response styles that can occur in a four-point rating scale. To do so, we hold the view that the IRTree model is a part of a larger modeling framework, called *explanatory item response modeling*, proposed by De Boeck and Wilson (2004). As discussed earlier, within this framework, the IRTree model can be formulated in GLNMM and specified as descriptive or explanatory models. In the study of response styles, the previous applications of the IRTree model were largely limited to a descriptive model, including only the indicator predictors. To our best knowledge, the explanatory IRTree model has never been considered. In the present study, we will employ both descriptive and explanatory IRTree models. Specifically, we will showcase how the explanatory IRTree model can help to inspect acquiescence and disacquiescence response styles, which have not been discussed in previous studies.

The remaining paper is organized as follows. We will first discuss the response styles investigated in this study. Next, we will showcase the application of the descriptive and explanatory IRTree models to study acquiescence, disacquiescence, and extreme response styles in a four-point Likert-type Rosenberg Self-esteem Scale.

3.1.1 Extreme, Acquiescence, and Disacquiescence Response Styles

The present study focuses on response styles in a four-point Likert-type rating scale consisting of response categories of ‘Strongly agree’ (SA), ‘Agree’ (A), ‘Disagree’ (D), and ‘Strongly disagree’ (SD). This type of scale not only asks the respondents to indicate their agreement or disagreement with the item statement, but also the extremity of their responses. In this scale, the respondents may have a tendency to use one of these response categories, leading

to the extreme, acquiescence, and disacquiescence response styles. In this section, I will review these response styles and explain how they will be investigated in this study.

Extreme Response Style. Extreme response style refers to a tendency to use the extreme response categories, irrelevant to the trait being measured (Baumgartner & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2013). This response style has been widely discussed because it can cause a bias in the measurement of the trait. For example, if the respondents tend to prefer or avoid the extreme categories, it can overestimate or underestimate their true level of the trait, and therefore their test scores may not properly represent the trait of interest. Because of this adverse effect, the extreme response style was regarded as a trait-irrelevant factor contaminating the measurement of the true trait.

Previous studies often assumed the extreme response style to be a unidimensional factor at the scale level (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010; Khorramdel & von Davier, 2014). However, this assumption may not be always true. For instance, in a rating scale having two extreme categories (e.g., 'Strongly agree' and 'Strongly disagree'), the behavior in choosing these two extreme categories might be quite different and point to two distinct extreme response styles. In the present study, we will test this assumption by examining the possibility of multidimensionality in the extreme response style. Moreover, we will evaluate where and how these extreme response styles occur among items. Previous studies often focused on evaluating the extreme response style at the scale level only. By inspecting the extreme response style at the item level, we will further provide rich and in-depth insight into the extreme response style.

Acquiescence and Disacquiescence Response Styles. The acquiescence response style describes a tendency, irrelevant to the trait being measured, to agree with the item statements, while the disacquiescence response style describes a tendency, irrelevant to the trait being measured, to disagree with the item statements (Baumgartner & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2013). Both the acquiescence and disacquiescence styles are a concern because their presence can induce measurement bias and contaminate the meaning of scores. Suppose that the A and SA are deployed to indicate a higher level of the trait (e.g., self-esteem), and D and SD to indicate a lower level of the trait. On such a scale, if the respondents tend to select SA or A, their test scores are very likely to be inflated. In contrast, the respondents' tendency to select D or SD can deflate their scale scores.

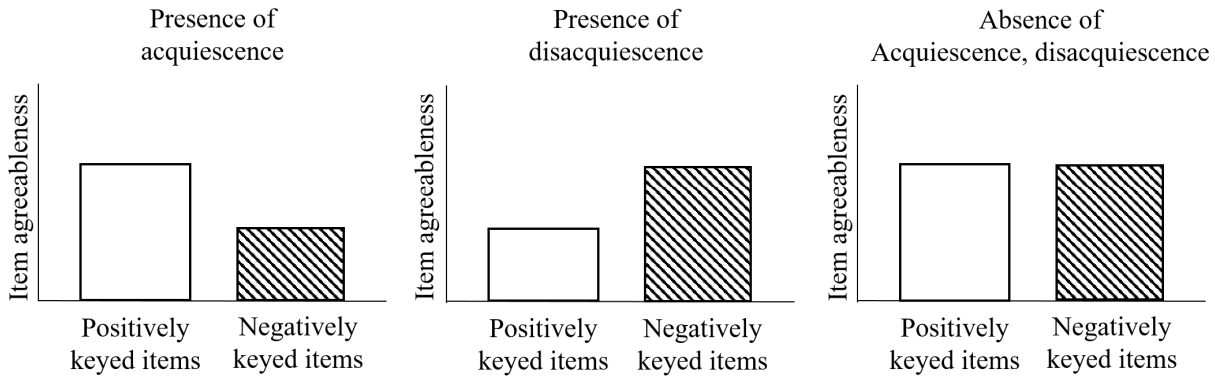
To minimize the effects of acquiescence and disacquiescence response styles, it is a common practice to construct a balanced scale where half of the items are positively keyed, and the other half are negatively keyed (Billiet & McClendon, 2000). Positively keyed items are phrased to represent a relatively high level of the trait by agreeing with the statements (e.g., "I am proud of myself." for measuring self-esteem), whereas negatively keyed items are phrased to represent a relatively high level of trait by disagreeing with the statements (e.g., "I certainly feel useless at times." for measuring self-esteem). Note that negatively keyed items can either be negatively worded grammatically, e.g., "I feel I do *not* have much to be proud of.", or positively worded grammatically, e.g., "I certainly feel *useless* at times." (Coleman, 2013). Hence, negatively keyed items are not always negatively worded items. It is believed that with an equal number of positively and negatively keyed items, the effects of the acquiescence and disacquiescence can be offset at the scale level.

In this paper, we attest that the data of a balanced scale contain useful information to detect the presence of acquiescence and disacquiescence response styles. Specifically, we will examine these two styles based on the pattern of *item agreeableness* statistics on a balanced scale. The item agreeableness statistic indicates how likely the response categories in an item, representing a high level of the trait, will be chosen. With mixed keyed items having four categories of SA, A, D, and SD, the item agreeableness can be defined as follows. The item agreeableness statistic of a positively keyed item indicates the likelihood of selecting the agree categories (i.e., A and SA) that reflect a high level of the trait. On the contrary, the item agreeableness statistic of a negatively keyed item indicates the likelihood of selecting the disagree categories (i.e., D and SD) that also reflect a high level of the trait.

The pattern of item agreeableness statistics for positively and negatively keyed items is informative for identifying acquiescence and disacquiescence response styles. If the respondents tend to choose the agree categories across items (i.e., acquiescence), this tendency will raise the item agreeableness levels of the positively keyed items, but lower the item agreeableness levels of the negatively keyed items. This will result in the agreeableness levels of the positively keyed items being higher than those of the negatively keyed items. On the contrary, if the respondents tend to choose the disagree categories across items (i.e., disacquiescence), this tendency will raise the item agreeableness levels of the negatively keyed items, but lower the item agreeableness levels of the positively keyed items. This will lead to the agreeableness levels of the negatively keyed items being higher than those of the positively keyed items. Following the same reasoning, when there is no acquiescence or disacquiescence, the item agreeableness levels will be similar between the positively keyed and negatively keyed items. These three scenarios

are presented in Figure 3.1. This study will examine the presence of acquiescence and disacquiescence based on this reasoning.

Figure 3.1 Presence of acquiescence and disacquiescence based on the pattern of item agreeableness levels of the positively and negatively keyed items.



In the following section, we will demonstrate the examination of extreme, acquiescence, and disacquiescence response styles using descriptive and explanatory IRTree models based on the real responses to the Rosenberg's Self-esteem Scale.

3.2 Method

3.2.1 Measure and Sample

The Rosenberg's Self-esteem Scale is a 10-item Likert-type rating scale widely used for measuring individuals' global self-worth. Each item has a statement about an individual's general feelings about oneself and requires respondents to indicate how strongly they agree or disagree with the statement. The scale is balanced with five positively keyed items and five negatively keyed items. In all ten items, there are four response categories of SA, A, D, and SD.

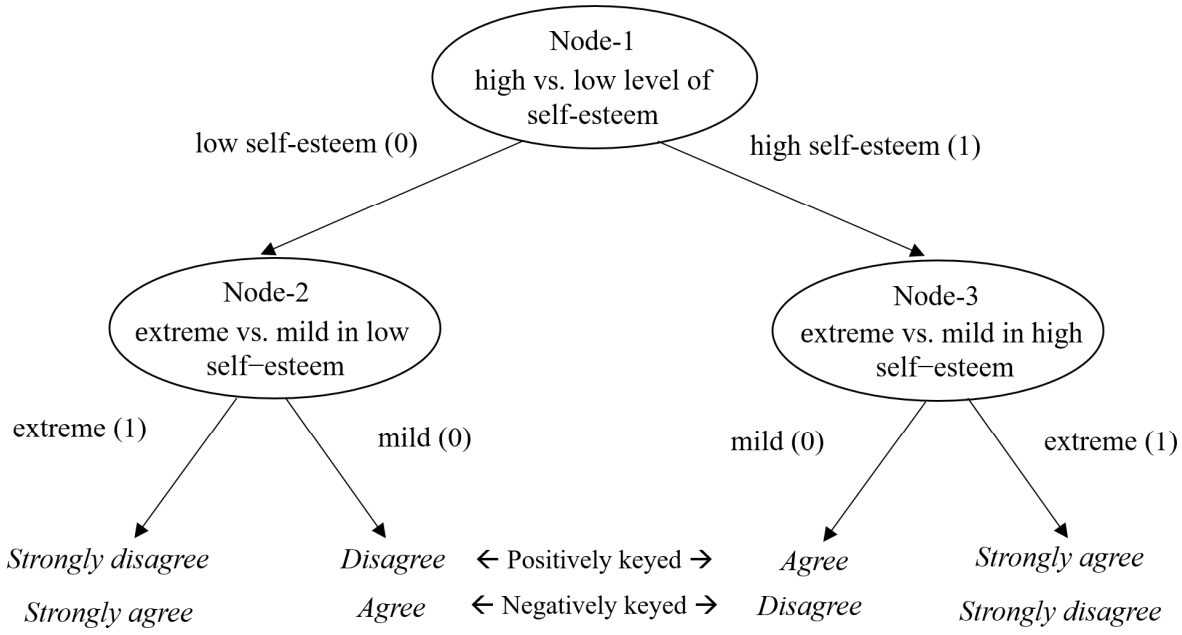
See Appendix A for the actual items. Data were retrieved from the 2005 Longitudinal Study of Generation in California (Silverstein & Bengtson, 2008). A total of 1,566 participants were included in the analysis. The majority of participants were female (56.8%), married (66%), and had a college or university degree (53.5%). The average age was 49.4 (SD_{age}=18.89, ranging from 16 to 98).

3.2.2 Model Specification

Tree Structure. To inspect extreme, acquiescence, and disacquiescence response styles by the IRTree model, we postulated the following decision process: (1) respondents determine whether they have positive feelings about themselves (i.e., a higher level of self-esteem) or negative feelings about themselves (i.e., a lower level of self-esteem), and then (2) they decide how strong their feelings are. The two-step decision process was described by a tree structure with three nodes in Figure 3.2. Node-1 represented whether the respondents chose the categories reflecting a high level of self-esteem (coded as 1) or the categories reflecting a low level of self-esteem (coded as 0). Thus, this node was referred to as *trait direction*. For the categories reflecting a lower level of self-esteem (i.e., when Node-1 branches out to 0), Node-2 represented whether the respondents chose the extreme category (coded as 1) or the mild category (coded as 0). This node was referred to as *extremity in low self-esteem direction*. For the categories reflecting a higher level of self-esteem (i.e., when Node-1 branches out to 1), Node-3 represented whether the respondents chose the extreme category (coded as 1) or the mild category (coded as 0). This node was referred to as *extremity in high self-esteem direction*. The participants' original choices of the four response categories were all recoded according to this tree structure. Note that response categories in the positively and negatively keyed items were recoded accordingly as the

categories can reflect a higher or lower level of self-esteem depending on the item keying direction.

Figure 3.2 Tree structure for detecting extreme, acquiescence, and disacquiescence response styles in the four-point Rosenberg's Self-esteem Scale.



Based on the tree structure, we specified descriptive and explanatory IRTree models to examine extreme, acquiescence, and disacquiescence response styles. Both the descriptive and explanatory models were specified as the generalized linear mixed models, hence, all models could be estimated by the *lme4* R package using the maximum likelihood estimation (Bates et al., 2015). In the following, we will explain these models.

Descriptive IRTree Model for Extreme Response Style. We first specified the descriptive IRTree model to examine the possibility of two distinct extreme response styles,

while controlling for self-esteem. The descriptive IRTree model included only the indicator predictors of person and item (i.e., person ID and item ID). For all three nodes in Figure 3.2, the effect of the person indicator variable was specified as random (latent variables), and the effect of the item indicator variable was specified as fixed (item parameters). Therefore, the logit of the probability of the outcome y^*_{ijn} of node Y^*_{ijn} was specified as follows,

$$logit\left(\pi\left(Y^*_{ijn} = y^*_{ijn}\right)\right) = \alpha1_{ij} * node1 + \alpha2_{ij} * node2 + \alpha3_{ij} * node3, \quad (3.1)$$

where

$$\alpha1_{ij} = \beta_{11} * Item1 + \beta_{12} * Item2 + \beta_{13} * Item3 + \beta_{14} * Item4 + \beta_{15} * Item5 + \beta_{16} * Item6 + \beta_{17} * Item7 + \beta_{18} * Item8 + \beta_{19} * Item9 + \beta_{110} * Item10 + \theta_{1i} \quad (3.1a)$$

$$\alpha2_{ij} = \beta_{21} * Item1 + \beta_{22} * Item2 + \beta_{23} * Item3 + \beta_{24} * Item4 + \beta_{25} * Item5 + \beta_{26} * Item6 + \beta_{27} * Item7 + \beta_{28} * Item8 + \beta_{29} * Item9 + \beta_{210} * Item10 + \theta_{2i} \quad (3.1b)$$

$$\alpha3_{ij} = \beta_{31} * Item1 + \beta_{32} * Item2 + \beta_{33} * Item3 + \beta_{34} * Item4 + \beta_{35} * Item5 + \beta_{36} * Item6 + \beta_{37} * Item7 + \beta_{38} * Item8 + \beta_{39} * Item9 + \beta_{310} * Item10 + \theta_{3i} \quad (3.1c)$$

The average logits of Node-1, Node-2, and Node-3 were indicated by the regression slopes $\alpha1_{ij}$, $\alpha2_{ij}$, and $\alpha3_{ij}$. Then, the average logit for each node was predicted by fixed effects of item indicators and random effect of person indicators. By plugging (3.1a), (3.1b), and (3.1c) into (3.1) and rearranging the equation, we can have a single equation shown below.

$$\begin{aligned} &logit\left(\pi\left(Y^*_{ijn} = y^*_{ijn}\right)\right) = \\ &(\beta_{11} * Item1 + \beta_{12} * Item2 + \beta_{13} * Item3 + \beta_{14} * Item4 + \beta_{15} * Item5 + \beta_{16} * Item6 + \beta_{17} * Item7 + \beta_{18} * Item8 + \beta_{19} * Item9 + \beta_{110} * Item10 + \theta_{1i}) * \boldsymbol{node1} \\ &+ (\beta_{21} * Item1 + \beta_{22} * Item2 + \beta_{23} * Item3 + \beta_{24} * Item4 + \beta_{25} * Item5 + \beta_{26} * Item6 + \beta_{27} * Item7 + \beta_{28} * Item8 + \beta_{29} * Item9 + \beta_{210} * Item10 + \theta_{2i}) * \boldsymbol{node2} \\ &+ (\beta_{31} * Item1 + \beta_{32} * Item2 + \beta_{33} * Item3 + \beta_{34} * Item4 + \beta_{35} * Item5 + \beta_{36} * Item6 + \beta_{37} * Item7 + \beta_{38} * Item8 + \beta_{39} * Item9 + \beta_{310} * Item10 + \theta_{3i}) * \boldsymbol{node3} \end{aligned} \quad (3.2)$$

This model resulted in three random effects for the three nodes. The first random effect represented individuals' levels of self-esteem, $\theta_{trait\ direction}$ for Node-1 (i.e., θ_{1i}). The second and third random effects represented individuals' extreme response styles in the low self-esteem direction, $\theta_{\text{extremity}(low\ trait)}$ for Node-2 (i.e., θ_{2i}) and in the high self-esteem direction, $\theta_{\text{extremity}(high\ trait)}$ for Node-3 (i.e., θ_{3i}). The two random effects for Node-2 and Node-3 allowed us to examine the presence of two extreme response styles in the opposite trait directions. To inspect extreme response styles in the scale, we compared the model with two distinct extreme styles to those of more constrained models – one with no extreme response styles at all, and the other with only one extreme response style regardless of the trait directions.

In addition to the three random effects, the model gave three sets of ten item parameters as the fixed effects, β_{jn}, one set for each of the three nodes (n = 1…3 for nodes, j = 1…10 for items). The set of ten item parameters for Node-1, $\beta_{trait\ direction:item}$ (i.e., β_{11}, β_{12}, ..., β_{110}), indicated the item agreeableness statistics, showing how likely the categories reflecting a high level of self-esteem would be chosen. The item parameters for Node-2 and Node-3 provided information about the extreme response styles at the *item* level. The set of ten item parameters for Node-2, $\beta_{extremity(low\ trait):item}$ (i.e., β_{21}, β_{22}, ..., β_{210}), indicated how likely the extreme category in the low self-esteem direction (SD in positively keyed items and SA in negatively keyed items) would be chosen. Likewise, the set of ten item parameters for Node-3, $\beta_{extremity(high\ trait):item}$ (i.e., β_{31}, β_{32}, ..., β_{310}), showed how likely the extreme categories in the high self-esteem direction (SA in positively keyed items and SD in negatively keyed items) would be chosen. These item parameters of Node-2 and Node-3 can be used to evaluate extreme response styles at the item level.

Explanatory IRTree Model for Acquiescence and Disacquiescence Response Styles.

To inspect the presence of acquiescence and disacquiescence response styles, the explanatory IRTree model was specified by including the person indicator variable and the item property variable (i.e., keying direction) as predictors. Because item property of keying direction was included as a predictor, this model was explanatory with respect to items. As for being fixed or random, the effect of the person indicator variable was treated as random (latent variables) and the effect of item property was treated as fixed (item parameters) for each node. It is worth noting that the item indicators were also included as random effects to control for the extra response variability due to item-by-item differences (e.g., due to contents). This specification is similar to the linear logistic test model *with error* as discussed by De Boeck (2008). Thus, the explanatory IRTree model was specified as follows:

$$logit\left(\pi\left(Y^*_{ijn} = y^*_{ijn}\right)\right) = \alpha 1_{ij} * node1 + \alpha 2_{ij} * node2 + \alpha 3_{ij} * node3, \quad (3.3)$$

where

$$\alpha 1_{ij} = \beta_{11} * Positively\ keyed + \beta_{12} * Negatively\ keyed + \theta_{1i} + u_{1j} \quad (3.3a)$$

$$\alpha 2_{ij} = \beta_{21} * Positively\ keyed + \beta_{22} * Negatively\ keyed + \theta_{2i} + u_{2j} \quad (3.3b)$$

$$\alpha 3_{ij} = \beta_{31} * Positively\ keyed + \beta_{32} * Negatively\ keyed + \theta_{3i} + u_{3j} \quad (3.3c)$$

The regression slopes $\alpha 1_{ij}$, $\alpha 2_{ij}$, and $\alpha 3_{ij}$ indicated the average logit of Node-1, Node-2, and Node-3. Then, for each node, the average logit was predicted by fixed effects of item keying direction and random effects of person and item indicators. By plugging (3.3a), (3.3b), and (3.3c) into (3.3) and rearranging the equation, we can have a single equation shown below.

$$logit\left(\pi\left(Y^*_{ijn} = y^*_{ijn}\right)\right) =$$

$$\left(\beta_{11} * Positively\ keyed + \beta_{12} * Negatively\ keyed + \theta_{1i} + u_{1j}\right) * \boldsymbol{node1} +$$

$$\left(\beta_{21} * Positively\ keyed + \beta_{22} * Negatively\ keyed + \theta_{2i} + u_{2j}\right) * \boldsymbol{node2} +$$

$$\left(\beta_{31} * Positively\ keyed + \beta_{32} * Negatively\ keyed + \theta_{3i} + u_{3j}\right) * \boldsymbol{node3} \quad (3.4)$$

The explanatory model resulted in three random effects of person indicators for the three nodes, representing individuals' levels of self-esteem and two extreme response styles, $\theta_{trait\ direction}$ (i.e., θ_{1i}), $\theta_{extremity(low\ trait)}$ (i.e., θ_{2i}), $\theta_{extremity(high\ trait)}$ (i.e., θ_{3i}) as well as three random effects of item indicators for the three nodes, representing item-by-item differences, $u_{trait\ direction}$ (i.e., u_{1j}), $u_{extremity(low\ trait)}$ (i.e., u_{2j}), and $u_{extremity(high\ trait)}$ (i.e., u_{3j}).

The model also yielded three sets of two item parameters as the fixed effects, β_{gn}, one set for each of the three nodes (n = 1…3 for node, g = 1 or 2 for item keying direction). The two item parameters for Node-1, $\beta_{trait\ direction:positively\ keyed}$ (i.e., β_{11}) and $\beta_{trait\ direction:negatively\ keyed}$ (i.e., β_{12}), indicated the overall item agreeableness levels for positively and negatively keyed items, respectively. The relative sizes of these two item agreeableness statistics revealed the presence of acquiescence and disacquiescence response styles (i.e., a higher agreeableness level for positively keyed items indicates acquiescence, and a higher agreeableness level for negatively keyed items indicates disacquiescence). The sets of item parameters for Node-2, $\beta_{extremity(low\ trait):positively\ keyed}$ (i.e., β_{21}) and $\beta_{extremity(low\ trait):negatively\ keyed}$ (i.e., β_{22}), and Node-3, $\beta_{extremity(high\ trait):positively\ keyed}$ (i.e., β_{31}) and $\beta_{extremity(high\ trait):negatively\ keyed}$ (i.e., β_{32}), indicated the potential effect of item keying direction on selecting the extreme categories. These two sets of parameters showed how likely the extreme categories would be chosen in the positively and negatively keyed items. This helped evaluate whether the extreme response styles would occur differently depending on the item keying direction.

The corresponding *lme4* R codes for descriptive and explanatory models were provided in Appendix D.

3.3 Results

3.3.1 Descriptive IRTree Model for Extreme Response Style

The descriptive model examined the extreme response style at the scale and item levels. At the scale level, the presence of the two distinct extreme response styles was evaluated by model fit comparisons. The results in Table 3.1 showed that the descriptive IRTree model with two extremity factors (Model 1c) fits noticeably better to the data, compared to the two other models – one specifying no extremity factor (1a in Table 3.1) and the other specifying a single extremity factor (1b in Table 3.1). This suggested the existence of two extreme response styles differentiated by the trait direction. The variance components (i.e., random effects) of these two extremity factors were noticeably greater than zero ($\theta_{extremity(low\ trait)}$ = 1.810 and $\theta_{extremity(high\ trait)}$ = 6.427) while controlling for the self-esteem factor ($\theta_{trait\ direction}$ = 4.461).

Table 3.1 Model fits of descriptive and explanatory IRTree models

Models	-2 LL	AIC	BIC
1a. Descriptive model with one trait factor and no ERS factors	23316.2	23378.0	23637.0
1b. Descriptive model with one trait factor and one ERS factor	22683.6	22730.0	22922.0
1c. Descriptive model with one trait factor and two ERS factors	22457.0	22529.0	22830.0
2. Explanatory model with one trait factor and two ERS factors	22620.8	22657.0	22807.0

Note. Trait represents self-esteem. ERS = extreme response style; -2 LL = -2 times log likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion. The lower values of -2 LL, AIC, and BIC indicate that the model fits better to the data.

To examine the two distinct extreme response styles at the item level, fixed effects of item indicators (βs) for Node-2 and Node-3 were evaluated (see Table 3.2). The estimates for Node-2 on the top panel of Table 3.2 showed where and how the extreme response style in the low-trait direction occurred in the items. For all items, the estimates were negative, suggesting that the extreme categories reflecting low self-esteem (SD in the positively keyed items and SA in the negatively keyed items) were less likely to be chosen, after controlling for self-esteem and item agreeableness. Likewise, the estimates for Node-3 on the bottom panel of Table 3.2 showed how the extreme response style in the high-trait direction (SD in the negatively keyed items and SA in the positively keyed items) occurred in the items. The results indicated no consistent pattern among the items after controlling for self-esteem and item agreeableness. The extreme categories in items 1, 2, 3, 5, and 10 were more likely to be chosen, whereas the extreme categories in items 4, 6, 7, 8, and 9 were less likely to be chosen.

Table 3.2 Estimates of fixed effects for items in the descriptive IRTree model

Node-2: extremity(low trait)	Item	$\beta_{extremity(low\ trait):item}$	SE	p
	Item 1	-0.877	0.443	0.048
	Item 2	-1.164	0.617	0.059
	Item 3	-1.646	0.356	<0.001
	Item 4	-2.302	0.383	<0.001
	Item 5	-1.657	0.368	<0.001
	Item 6	-3.173	0.462	<0.001
	Item 7	-2.676	0.377	<0.001
	Item 8	-1.901	0.204	<0.001
	Item 9	-2.751	0.294	<0.001
	Item 10	-2.424	0.326	<0.001
Node-3: extremity(high trait)	Item	$\beta_{extremity(high\ trait):item}$	SE	p
	Item 1	0.925	0.099	<0.001
	Item 2	0.228	0.097	0.019
	Item 3	0.966	0.100	<0.001
	Item 4	-0.682	0.099	<0.001
	Item 5	0.499	0.099	<0.001
	Item 6	-1.143	0.101	<0.001
	Item 7	-1.151	0.101	<0.001
	Item 8	-0.991	0.107	<0.001
	Item 9	-0.388	0.102	<0.001
	Item 10	0.647	0.101	<0.001

Note. Negatively keyed items are underlined.

3.3.2 Explanatory IRTree Model for Acquiescence and Disacquiescence Response Styles

The explanatory IRTree model examined the presence of acquiescence and disacquiescence styles. This model included the self-esteem factor and two extremity factors ($\theta_{trait\ direction}$= 4.253, $\theta_{extremity(low\ trait)}$ = 1.833, $\theta_{extremity(high\ trait)}$ = 6.337) as well as item-by-item differences ($u_{trait\ direction}$= 0.964, $u_{extremity(low\ trait)}$= 0.345, and $u_{extremity(high\ trait)}$= 0.579). These effects were controlled for while examining the presence of acquiescence and disacquiescence response styles.

To evaluate the presence of acquiescence and disacquiescence styles, the fixed effects of item keying direction for Node-1 (βs), corresponding to the item agreeableness, were examined in Table 3.3. The results showed that the item agreeableness levels were noticeably higher for the positively keyed items than for the negatively keyed items, suggesting the presence of acquiescence response style. Furthermore, the fixed effects of item keying for Nodes-2 and -3 were examined to evaluate the respondents' uses of extreme categories in positively and negatively keyed items. The results showed that the respondents tended to avoid using the extreme categories reflecting a low level of trait for both item keying directions (see the negative estimates for Node-2 in Table 3.3). The extent of the avoidance was fairly comparable between positively and negatively keyed items. By contrast, there was no significant preference or avoidance of the extreme categories reflecting a high level of trait for both positively and negatively keyed items (see the estimates for Node- 3 in Table 3.3).

Table 3.3 Estimates of fixed effects for item keying in the explanatory IRTree model

Node 1: trait direction	Item Keying	$\beta_{trait\ direction:keying}$	SE	p
	Positively keyed	4.756	0.447	<0.001
	Negatively keyed	3.193	0.440	<0.001
Node-2: extremity(low trait)	Item Keying	$\beta_{extremity(low\ trait):keying}$	SE	p
	Positively keyed	-2.075	0.383	<0.001
	Negatively keyed	-2.015	0.343	<0.001
Node-3: extremity(high trait)	Item Keying	$\beta_{extremity(high\ trait):keying}$	SE	p
	Positively keyed	-0.363	0.342	0.289
	Negatively keyed	0.143	0.342	0.676

3.4 Discussion

This study extended the applicability of an IRTree model in the study of response styles. Specifically, we showcased how the IRTree model, either descriptive or explanatory, can be stipulated from the vantage point of a generalized linear mixed model, under the explanatory item response modeling framework. As a demonstration, the extreme, acquiescence, and disacquiescence response styles were examined based on the responses to the Rosenberg's Self-esteem Scale. Our findings suggested the existence of two distinct extreme response styles in the low and high trait directions. The two extremity styles were examined at the item level as well. In all items, people tended to avoid the extreme categories that reflect a low level of self-esteem. The extent of the avoidance was fairly comparable between positively and negatively keyed items. In contrast, only for some items but not all, people tended to avoid the extreme categories that reflect a high level of self-esteem. No significant preference/avoidance of the extreme

categories was found for both positively and negatively keyed items. Moreover, our findings pointed to the acquiescence response style, but not the disacquiescence response style, when responding to the Rosenberg's Self-esteem Scale.

The present study makes new contributions in several ways. First, it explored the possibility of two distinct extreme response styles. Not only that, it evaluated the two extreme response styles in depth by looking into where and how the response styles occurred among the items. Second, we introduced the explanatory IRTree model, which has not been considered in previous studies of response styles. Our study demonstrated the explanatory IRTree model by incorporating item keying direction as a predictor and showed how the acquiescence and disacquiescence response styles can be indirectly detected without entailing external measures. Lastly, this article showcased the versatility of the IRTree models in the study of response styles, when conceived under the explanatory item response modeling framework. The IRTree model can be specified by combining different predictors, either as an indicator or a property, for person, item, and response category. This permits researchers to build a variety of models tailored to their own research hypotheses.

We would like to point out that the findings on response styles have several implications on the Rosenberg's Self-esteem Scale. First, the results of the extreme response styles indicate that the scale scores can be overestimated. That is, the respondents tended to consistently avoid the extreme categories that reflect a low level of self-esteem. For example, when respondents disagree with the statement 'I am proud of myself', they tended to avoid extreme answers (i.e., Strongly disagree). This tendency can result in respondents having higher scale scores than their true level of self-esteem. These overestimated scores can lead to a biased measure of self-esteem. Moreover, when it is used in research, it can deflate/inflate the correlations of self-esteem scores

with other measures. Hence, researchers should be aware of this issue when interpreting and using the Rosenberg Self-esteem Scale scores. Researchers may also consider statistical correction to control for the impact of extreme response styles.

Second, the results of the acquiescence response style suggest that the deployment of a balanced scale can be beneficial to the Rosenberg's Self-esteem Scale. This study showed that the respondents tended to agree with the statements (i.e., acquiescence) when responding to the questions. If the scale was not balanced with positively and negatively keyed items, the scale scores could be very likely to be inflated due to this tendency. As mentioned, the inflated scale scores can cause measurement biases and lead to dubious correlations with other measures. By using a balanced scale, the effect of the acquiescence response style can be mitigated.

The investigation of response styles can also provide an additional way of evaluating the validity of the Rosenberg's Self-esteem Scale. Validity is defined as "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (AERA et al., 2014). Thus, validation is an ongoing process by accumulating various sources of evidence. For the Rosenberg's Self-esteem Scale, validity has been evaluated based on evidence such as its factor structure (e.g., Hyland et al., 2014; Sinclair et al., 2010; Vermillion & Dodder, 2007) and relations with other measures (e.g., Bagley & Mallick, 2001; Hagborg, 1993). Evaluating response styles can provide an additional way of validation because it helps evaluate measurement bias that can affect the meaning of the scores. In the current study, we detected the acquiescence and extreme response styles, showing that the scores of the Rosenberg's Self-esteem Scale may be obfuscated by them. The presence of these response styles could undermine the validity of the score meaning. In the validation, these findings can help evaluate the interpretation of the scale score as extra evidence.

It is reasonable to conjecture that the response styles identified in this study travel well to other measures of global self-esteem that have the same response format (i.e., four-point Likert-type rating scale comprising positively and negatively keyed items) and for similar population. However, these findings should not be generalized to measures of other constructs in different response formats, and/or for specific populations such as children, seniors, and clinical populations without further empirical evidence. We encourage future studies to verify these specific generalizations.

As a caveat, the findings of the present study were based only on a single dataset, hence further cross-validation is needed. In several previous studies, response styles were investigated based on multiple sources of evidence. For example, Zettler et al. (2016) used the self- and observer-report measures of personality traits for the same individuals to detect response styles. They emphasized that the consistency in findings from cross-source data is essential to verify the presence of response styles. In other reports, external measures were employed to verify the response styles (e.g., Plieninger & Meiser, 2014). Our findings were not based on cross-source data, nor were they compared to any external criteria. In this respect, we encourage future investigations to cross-validate the current findings.

Finally, as being showcased, the IRTree model permits researchers to disentangle response styles from the substantive trait by modeling the trait-relevant and trait-irrelevant factors as separate nodes in the tree structure. This feature enables researchers to differentiate those respondents who carry a response style from those who do not. For example, following the tree structure in Figure 3.2, it is possible to distinguish a respondent, say Mary, with an exceedingly high level of self-esteem but *no* extreme response styles from another respondent,

say John, with a high level of self-esteem *and* extreme response styles (say, preferring extreme responses in both trait directions). Mary and John could have the same responses to an item (SD or SA), hence follow the same path in the tree structure. However, Mary would be estimated to have a higher score on the trait factor, compared to John. On the contrary, John would be estimated to have a higher score on the extreme response style factors, compared to Mary. In this respect, the IRTree model allows researchers to disentangle response styles from the substantive trait.

Chapter 4: Examining Person and Item Characteristics Associated with Not-reached and Omitted Responses Using Item Response Tree Models

4.1 Introduction

Large-scale educational assessments play a key role to evaluate students' progress in learning. The outcomes of these tests are a crucial source of information on teaching, performance development, and educational policy (Cresswell et al., 2015). Well-known examples include the Progress in International Reading Literacy Study, the Trends in International Mathematics and Science Study, the Programme for International Student Assessment, and the National Assessment of Educational Progress. In these educational assessments, it is not uncommon that test-takers omit the questions or do not complete the test. Such behaviors are generally referred to as item nonresponse (De Leeuw et al., 2003; Groves, 1989; Huisman, 1999; Köhler et al., 2017).

Large-scale educational assessments often discern two types of item nonresponse: omitted and not-reached. For example, the Progress in International Reading Literacy Study (PIRLS) defines omitted and not-reached responses as follows (Martin et al., 2017). A nonresponse is defined as an *omitted response* when the test-takers left the item blank, or the response was uninterpretable or out-of-range. In contrast, a nonresponse is defined as a *not-reached response* when the test-takers did not attempt the items due to a lack of time. In other words, the not-reached response is flagged when test-takers fail to complete the test. This type of nonresponse is featured by the pattern of sequential missing near the end of the test. One

example of such a pattern is (1, 0, 1, 1, 0, 9, 9, 9, 9, 9), where 1 and 0 denote correct and incorrect answers respectively, and 9 is a missing response.

It is critical to understand how these two types of item nonresponse occur because they can cause serious biases in the assessment of student performance and test fairness (Köhler et al., 2017; Mislevy & Wu, 1996; Rose, 2013). Suppose that test-takers tend to omit some types of items (e.g., constructed-response items) more frequently. This can lead to a systematic failure in evaluating the skills/knowledge measured by these types of items. In such a case, the resultant test score may not properly reflect one's level of ability. Meanwhile, a particular group of test-takers (e.g., boys) may tend to omit the items more frequently. This can systematically underestimate their ability because most educational testing programs treat an item nonresponse as an incorrect answer in the scoring process (Martin et al., 2016, 2017). Thus, those who have a higher rate of item nonresponse tend to have lower test scores and could be mistakenly underestimated. This makes it difficult to have a fair comparison to other test-takers.

Despite its negative impacts, item nonresponse behaviors were not given much attention in the psychometric literature. To date, only several studies have investigated how item nonresponse occurs (e.g., Di Chiacchio et al., 2016; Köhler et al., 2015; Koretz et al., 1993; Matters & Burnett, 1999, 2003; Okumura, 2014). Most of these previous studies examined item nonresponse in large-scale educational assessments such as the Programme for International Student Assessment (PISA), the National Assessment of Educational Progress (NAEP), the National Educational Panel Study (NEPS), and the Queensland Core Skills (QCS) Test. Okumura (2014) examined the omitted response in the PISA reading assessment among Japanese students. It reported that the omitted response was more likely to occur for girls and open-ended items (vs. boys and multiple-choice items). Girls also tended to omit multiple-choice items more

frequently than boys, whereas the reversed pattern was found for open-ended items (i.e., boys were more likely to omit than girls). This study also identified other related factors such as social-cultural economic status, enjoyment of reading, and teachers' stimulation of reading engagement. More recently, Di Chiacchio et al. (2016) investigated both not-reached and omitted responses in the PISA science assessment among Italian students. They found that the not-reached response was more prevalent for boys and test-takers having lower self-efficacy and greater enjoyment and interest in science. Meanwhile, the omitted response occurred more frequently for test-takers having lower self-efficacy and limited enjoyment in science.

Koretz et al. (1993) investigated both not-reached and omitted responses in the NAEP mathematics assessment among U.S. students. The authors found that open-ended items and minority ethnic groups were more likely to have a higher rate of omitted response (vs. multiple-choice items and Caucasian). They also reported a weak relationship between item format and not-reached response, implying that not completing the test is less affected by item format. With the NEPS, Köhler et al. (2015) examined item nonresponse behaviors in multiple competence domains of information and communication technologies, science, mathematics, and reading among German students. They reported that test-takers attending lower secondary schools[1] and having a migration background were more likely to not reach and omit their responses than their counterparts (i.e., upper secondary schools, no migration).

Matters and Burnett (1999) examined the Queensland Core Skills (QCS) Test – a statewide achievement test evaluating the common curriculum in Australia. They examined the

[1] The lower secondary schools are loosely equivalent to junior high schools; the upper secondary schools are equivalent to senior high schools (UNESCO Institute for Statistics, 2012).

omitting response behavior separately for short-response and multiple-choice items. For short-response items, they found that boys and students from government schools were more likely to omit responses than their counterparts (i.e., girls and those from non-government schools). They also found that girls from single-sex schools omitted less often than students from coeducational schools. The same pattern was found for multiple-choice items. In their follow-up work, Matters and Burnett (2003) further reported that test-takers with a lower level of academic self-efficacy, self-estimate of ability, and motivation were more likely to omit the items.

Worth noting is that these previous studies used different statistical techniques. The earlier works relied mostly on descriptive statistics (e.g., rates of omitted response) or correlations (Koretz et al., 1993; Matters & Burnett, 1999, 2003). Later, Di Chiacchio et al. (2016) used cluster analysis to identify groups of test-takers based on their patterns of item nonresponse (lower omitter, leaver, and jumper). They then examined how the clusters were associated with the test takers' characteristics. Not until recently have some studies adopted advanced item response modeling techniques (Köhler et al., 2015; Okumura, 2014) and examined the reasons for their occurrence.

As a recent method, the item response tree (IRTree) model was introduced as a promising tool for modeling a variety of response behaviors, including item nonresponse. The IRTree model enables researchers to analyze item responses by postulating a decision process via a tree structure. Taking this feature, several studies have shown that the IRTree model is not only useful for analyzing nonresponse behaviors but also for inspecting the effects of item and test-taker characteristics on them. De Boeck and Partchev (2012) were the first to suggest using the IRTree model for examining item nonresponse. They proposed to treat the omitted response as a response category and hypothesize the process leading to this category in a tree structure. Later

on, Okumura (2014) followed this idea with real data to inspect the effects of the various item and test-taker characteristics, such as item format, sex, and enjoyment of reading. Okumura, however, only examined the omitted response. In a more recent work, Debeer et al., (2017) hypothesized several plausible decision processes for both not-reached and omitted responses and tested them with both simulated and real data. However, their focus was not on evaluating the effects of person and item characteristics.

The present study used the IRTree model to explain the occurrence of *both* not-reached and omitted response behaviors in the PIRLS reading test. To do so, our explication of the IRTree model will take the lenses of the *explanatory item response modeling framework* (De Boeck & Wilson, 2004). By adopting this framework, we can easily extend the standard IRTree model to the explanatory IRTree model as a form of a generalized linear and nonlinear mixed model (GLNMM). In this study, we will show how the explanatory IRTree model can help investigate various person and item characteristics related to both not-reached and omitted responses.

4.1.1 Item Nonresponse in PIRLS

The PIRLS has been widely used to evaluate reading literacy worldwide. Yet, there is limited information on item nonresponse behaviors when taking this test. In particular, the person and item characteristics that may lead to systematic differences in item nonresponse have not been examined. To fill this gap, this study will inspect the following four characteristics that can have practical ramifications on the PIRLS test scores: gender, test language, item location, and item format.

First, we will examine potential gender differences in both not-reached and omitted responses. Gender was previously reported as a relevant factor in item nonresponse (Di Chiacchio et al., 2016; Matters & Burnett, 1999; Okumura, 2014). It is reasonable to believe that gender may also impact nonresponse in a large-scale reading test like the PIRLS. This can help identify whether one gender is disadvantaged due to item nonresponse. This study will also investigate the effect of item format. Many educational tests, including the PIRLS, use two item formats, multiple-choice and constructed-response (a.k.a. open-ended items). Previous studies observed that the constructed-response item is more prone to omission (Koretz et al., 1993; Okumura, 2014). This can cause a systematic failure in assessing the specific skills/knowledge measured by this format. This study will test whether this would occur in the PIRLS as well.

In this study, we will further examine two new factors, test language and item location, which have not been studied in previous research. The test language is an important factor in international testing like the PIRLS because it is translated into numerous languages. Hence, it is critical to ensure the comparability between different versions of translated tests. If item nonresponse is more prevalent for students taking either version, it can cause one test-language group to be disadvantaged and lead to a serious bias in performance comparison. Item location is also an important design factor in educational testing. Previous studies pointed out that test-takers' responses can be different depending on the order that items appear in the test (Hambleton & Traub, 1974; Laffitte, 1984; Lavrakas, 2008a; Zwick, 1991). Thus, we can anticipate that the location of where the items are placed in the test will affect the occurrence of nonresponse as well. If this indeed happens, the test scores may fail to reflect the particular skills/knowledge that the missed questions intend to assess.

In the next section, we will describe the PIRLS data and the specification of the explanatory IRTree model to examine item nonresponse behaviors.

4.2 Method

4.2.1 Measure and Sample

The PIRLS 2016 is an international assessment that aims to measure student learning in reading. It asks the students to answer multiple sets of items, each regarding a reading passage, in order to assess their comprehension of the content. The questions are either in the format of multiple-choice or constructed-response. For missing responses, the PIRLS indicated them either as not-reached or omitted. For valid responses to the questions, this study scored them either as correct (code = 1) or incorrect (code = 0) for both types of items.[2]

This study analyzed the data for 18,245 fourth-grade students in Canada. The dataset included the responses to the 181 items in all 13 booklets. The included person and item characteristics were: gender (boy = 50.1%, girl = 49.9%), test language (English = 67%, French = 33%), item format (multiple-choice = 47.5%, construct response = 52.5%), and item location. For item location, because the number of items was not the same for different booklets, we calculated a value indicating an item's *relative* position in its given booklet. That is, each item's sequential position in the test was divided by the total number of items in that booklet. This made

[2] Originally, the PIRLS scored multiple-choice items as either correct or incorrect, while constructed-response items were scored as incorrect, partially correct, and correct. For consistency purpose, this study re-scored the constructed-response items into binary scores by treating the incorrect and partially correct responses as incorrect.

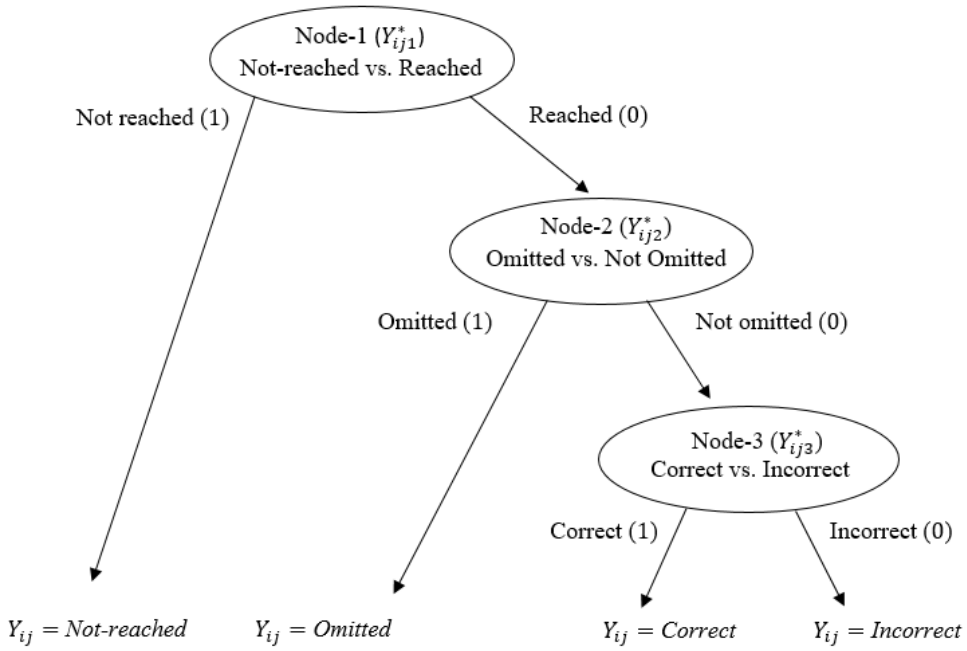
the item location comparable across the booklets. We then standardized the values of relative item location so that it has a mean of zero and standard deviation of one. Hence, conceptually, an item with a large positive value means its relative location is far behind the middle location (i.e., closer to the end of the test). The rest of the three predictors were dummy coded: gender (girl = 0, boy = 1), test language (English = 0, French = 1), and item format (multiple-choice = 0, constructed-response = 1). The extent of multicollinearity between the predictors was tested by the variance inflation factor (VIF). The VIF values of the predictors ranged from 1.04 to 2.27, indicating that the predictors were not highly correlated with each other (Cohen et al., 2003).

In the analysis, we split the entire dataset into two random halves. The specified IRTree model (to be explained) was fitted to the first half. The second half was used to cross-validate the findings by checking whether we can replicate the results from the first half.

4.2.2 Model Specification

Tree structure. This study specified a decision process that simultaneously models not-reached and omitted responses, shown in Figure 4.1, as suggested by Debeer et al. (2017). This process consisted of three decision steps, leading to four response categories: not reached, omitted, correct, and incorrect. Node-1 (Y^*_{ij1}) represented the first step regarding whether the students attempted the item (i.e., reached; branch coded as 0) or did not attempt the item (i.e., not reached; branch coded as 1). For students who attempted the item, as the second step, Node-2 (Y^*_{ij2}) specified a decision about whether they omitted the items (branch coded as 1) or did not omit the items (branch coded as 0). If the students provided a valid response (i.e., did not omit), the last step in Node-3 (Y^*_{ij3}) specified whether the response was correct (branch coded as 1) or incorrect (branch coded as 0).

Figure 4.1 Tree structure for not-reached and omitted responses.



Following this decision process, we recoded the original four observed responses (Y_{ij}) for person *i* to item *j* into a set of node outcomes, shown as a mapping matrix in Table 4.1. The not-reach response was recoded as (1, NA, NA) for the three nodes; the omitted response was coded as (0, 1, NA); the correct response was coded as (0, 0, 1); the incorrect response was coded as (0, 0, 0). The NA indicates that a decision step was not applicable. For example, if a question was not reached (Node-1), it was not applicable to consider whether it was omitted (Node-2) or answered correctly (Node-3), and therefore these nodes were coded as NA. Appendix B shows a sample of the recoded data according to this mapping matrix.

Table 4.1 Mapping matrix for the IRTree model

Responses (Y_{ij})	Node-1 (Y^*_{ij1})	Node-2 (Y^*_{ij2})	Node-3 (Y^*_{ij3})
Not-reached	1	NA	NA
Omitted	0	1	NA
Correct	0	0	1
Incorrect	0	0	0

Note. NA denotes not applicable.

Given the decision process, we specified the explanatory IRTree model to investigate the effects of person and item characteristics on the not-reached and omitted responses. The explanatory IRTree model, as a generalized linear mixed model, was estimated by the *lme4* R package using maximum likelihood estimation (Bates et al., 2015). In the following, we will explain its specification in detail.

Explanatory IRTree model. We specified the explanatory IRTree model by including item location, item format, gender, and test language as predictors. To inspect not-reached response, we specified item location, gender, and test language as having fixed effects in Node-1. For the omitted response, we included an additional factor, item format, as having a fixed effect for Node-2. Note that the item indicators were specified as having random effects on Node-1 and Node-2 so as to account for unexplainable variability due to item-by-item differences. This specification is similar to the linear logistic test model *with error* as discussed by De Boeck (2008). Moreover, we also controlled for student ability and item difficulty. It has been known that student ability and item difficulty are related to not-reached and omitted responses (Di Chiacchio et al., 2016; Köhler et al., 2015). That is, not-reached and omitted responses are more

likely to occur when test-takers have a lower level of ability and when items are more difficult. Therefore, these factors can be confounders when the major research interest is the effects of item/person properties. Thus, to control for student ability and item difficulty, we included person and item indicators as having random effects on Node-3. Hence, the logit of the probability of the outcome for node Y^*_{ijn} was specified as follows.

$$logit\left(\pi\left(Y^*_{ijn} = y^*_{ijn}\right)\right) = \alpha 1_{ij} * node1 + \alpha 2_{ij} * node2 + \alpha 3_{ij} * node3, \quad (4.1)$$

where

$$\alpha 1_{ij} = \beta_{10} + \beta_{11} * ItemLoc + \beta_{12} * Gender + \beta_{13} * TestLang + u_{1j} \quad (4.1a)$$

$$\alpha 2_{ij} = \beta_{20} + \beta_{21} * ItemLoc + \beta_{22} * Gender + \beta_{23} * TestLang + \beta_{24} * ItemForm + u_{2j} \quad (4.1b)$$

$$\alpha 3_{ij} = \beta_{30} + \theta_{3i} + u_{3j} \quad (4.1c)$$

The average logits of Node-1, Node-2, and Node-3 were specified by $\alpha 1_{ij}$, $\alpha 2_{ij}$, and $\alpha 3_{ij}$. Then, for each node, the average logit was predicted by fixed effects of person and item characteristics and random effects. By plugging (4.1a), (4.1b), and (4.1c) into (4.1) and rearranging the equation, the explanatory IRTree model can be presented as follows.

$$\begin{aligned} &logit\left(\pi\left(Y^*_{ijn} = y^*_{ijn}\right)\right) = \\ &\left(\beta_{10} + \beta_{11} * ItemLoc + \beta_{12} * Gender + \beta_{13} * TestLang + u_{1j}\right) * \boldsymbol{node1} + \\ &\left(\beta_{20} + \beta_{21} * ItemLoc + \beta_{22} * Gender + \beta_{23} * TestLang + \beta_{24} * ItemForm + u_{2j}\right) * \boldsymbol{node2} + \\ &\left(\beta_{30} + \theta_{3i} + u_{3j}\right) * \boldsymbol{node3} \end{aligned} \quad (4.2)$$

First, the model included three conditional means, one for each node, β_{Node-1} (i.e., β_{10}), β_{Node-2} (i.e., β_{20}), and β_{Node-3} (i.e., β_{30}). These parameters indicated the average logits, hence the probabilities, of not-reached, omitted, and correct responses.

Second, the IRTree model led to a set of fixed effects of person and item properties on Node-1 and Node-2, separately. For Node-1 (not-reached), the model yielded three fixed effects: (1) the parameter $\beta_{Node-1:item\ location}$ (i.e., β_{11}), evaluating whether the not-reached response occurred differently depending on the item location, (2) the parameter $\beta_{Node-1:gender}$ (i.e., β_{12}), examining the gender difference in the not-reached response, and (3) the parameter $\beta_{Node-1:test\ language}$ (i.e., β_{13}), evaluating the difference between the English and the French versions in the not-reached response.

For Node-2, the first three predictors were the same, hence their fixed effect estimates, $\beta_{Node-2:item\ location}$ (i.e., β_{21}), $\beta_{Node-2:gender}$ (i.e., β_{22}), and $\beta_{Node-2:test\ language}$ (i.e., β_{23}) had similar meaning except that omitting behavior was being explained. The fourth fixed effect, $\beta_{Node-2:item\ format}$ (i.e., β_{24}), was the difference between the constructed-response and multiple-choice items in the omitted response.

Finally, as controls, this model yielded two random effects for Node-3, representing student ability, θ_{Node-3} (i.e., θ_{3i}), and item difficulty, u_{Node-3} (i.e., u_{3j}). In addition, two random effects of item indicators for Node-1 and Node-2, u_{Node-1} (i.e., u_{1j}) and u_{Node-2} (i.e., u_{2j}), were included, representing additional variability due to items in not-reached and omitted responses.

The corresponding *lme4* R code for the specified model was provided in Appendix D.

4.3 Results

The explanatory model showed a better fit compared to the descriptive model (IRTree model with no person and item property variables) as shown by its lower model fit indices

reported in Table 4.2. Because the descriptive model was nested within the explanatory model, the Chi-square difference test was further conducted to examine the model fit. The $\Delta X^2(7) = 1934.9$, p <.001, showing that there was a significant improvement in fit to the data from the descriptive model to the explanatory model.

Table 4.2 Model fits of descriptive and explanatory IRTree models

Models	-2 LL	AIC	BIC
1. Descriptive IRTree model	350941.9	350961.9	351077.7
2. Explanatory IRTree model	349007.0	349041.0	349237.8

Note. -2 LL = -2 times log likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion. A lower value of -2 LL, AIC, and BIC indicates that the model fits better to the data.

The conditional means of the three nodes were estimated ($\beta_{Node-1} = -6.668$, $\beta_{Node-2} = -4.639$, and $\beta_{Node-3} = 0.842$). This showed that, on average, the chance of getting the correct answer was way higher than the other two nonresponse behaviors.

It also estimated the variances of ability and item difficulty as random effects ($\theta_{Node-3} = 1.172$ and $u_{Node-3} = 2.084$). The additional item variabilities for Node-1 and Node-2 were estimated as well ($u_{Node-1} = 0.938$ and $u_{Node-2} = 0.520$). These random effects were controlled for while examining the fixed effects of person and item properties.

Of the most interest was the fixed effects of person and item properties, reported in Table 4.3. For Node-1, the results showed that item location and test language were significantly associated with the not-reached response. For item location, the positive estimate ($\beta_{Node-1:item\ location} = 1.217$) indicated that the items located later in the test were more likely to be not reached (i.e., not attempted). The odds of not being reached increased by 3.38 times for

every one standard deviation increase in the relative item location. As for test language, the estimate ($\beta_{Node-1:test\ language} = 0.841$) showed a significant difference between the two test languages. The odds of having a not-reached response were 2.32 times higher for the students who took the French version than for those taking the English version. This indicated that the French version students were more likely to fail to complete the test than the English version students. The estimate of gender ($\beta_{Node-1:gender} = 0.064$), however, was not significant.

Table 4.3 The fixed effects of the person and item characteristics in the explanatory IRTree model

Node-1 (not-reached)		*β*	*SE*	*p*	*Odds ratio*
	Item location	1.217	0.087	<.001	3.38
	Gender	0.064	0.050	0.196	1.07
	Test language	0.841	0.050	<.001	2.32
Node-2 (omitted)		*β*	*SE*	*p*	*Odds ratio*
	Item location	0.266	0.045	<.001	1.30
	Item format	0.989	0.086	<.001	2.69
	Gender	0.129	0.021	<.001	1.14
	Test language	0.746	0.021	<.001	2.11

For Node-2, the fixed effects of item location, item format, gender, and test language on the omitted response were all significant. The positive estimate of item location ($\beta_{Node-2:item\ location} = 0.266$) suggested that the items located later in the test were more likely to be omitted. The odds of being omitted increased by 1.30 times for every one standard deviation increase in the relative item location. For item format, the estimate

($\beta_{Node-2:item\ format} = 0.989$) indicated that the constructed-response items were more likely to be omitted than multiple-choice items. The odds were 2.69 times higher for the constructed-response items than for the multiple-choice items. In terms of gender, the estimate ($\beta_{Node-2:gender} = 0.129$) showed a significant difference in the omitted response. The odds of omitting the items were 1.14 times higher for boys compared to girls. Lastly, the estimate of test language ($\beta_{Node-2:test\ language} = 0.746$) showed that students in the French version were more likely to omit the items than those in the English version. The odds were 2.11 times higher for the French version takers than their counterparts.

The results were cross-validated and we found similar results with the second random half of the data (see Appendix C).

4.4 Discussion

This study introduced the explanatory IRTree model to examine the person and item characteristics related to item nonresponse behaviors in the PIRLS reading assessment among grade four Canadian students. With regards to person characteristics, gender and test language were found to be significant factors. Specifically, students taking the French version failed to complete the test (i.e., not reached) and omitted the items more frequently than those taking the English version. In comparing between boys and girls, boys were more likely to omit questions; however, the difference was small (odds ratio = 1.14). As for item properties, item location and item format were found to be significant factors. The closer an item was to the end of the test, the more likely the not-reached and omitted responses would occur. The omitted response was also more likely to happen to constructed-response items.

The present study contributes to the literature of item nonresponse in several ways. First and foremost, this study broadened the utility of the IRTree model. Although the IRTree model has been applied to study item nonresponse and identify relevant factors (Debeer et al., 2017; Okumura, 2014), previous studies have not studied the effects of properties on *both* not-reached and omitted responses in one single integrated IRTree model. This study showcased how this could be done through the lenses of the explanatory item response modeling. Moreover, this study shed light on item nonresponse behaviors when taking the reading test of the PIRLS. Despite being used and researched worldwide, there were limited reports on item nonresponse. This study provided evidence on the effects of gender, test language, item location, and item format, which help to understand nonresponse behaviors and their undesirable effects on the PIRLS.

The findings of this study are related to the literature in a number of aspects. First, the findings of gender differences were somewhat inconsistent with the extant literature. Inconsistent with Di Chiacchio et al. (2016) that reported Italian boys having a higher rate of not-reached response in the PISA science test, this study showed no significant gender difference in the not-reached response. Although the difference was small, this study found that boys tended to omit the items more frequently. This is opposite to the report by Okumura (2014) where Japanese girls were found more likely to omit the items in the PISA reading test. These inconsistencies suggest that the gender differences in nonresponse may occur differently depending on the subject (e.g., reading, math) and/or samples (e.g., Canadian, Italian, and Japanese). Perhaps, nonresponse behaviors are specific to different subjects and educational cultures. Given the scarcity of literature on nonresponse, further investigation is needed to illuminate these inconsistencies.

The observed gender difference in this study raises some concerns about fairness in performance comparison. Many studies reported that girls outperformed boys in reading (Chiu & McBride-Chang, 2006; Logan & Johnston, 2010; Mullis et al., 2003, 2007). Likewise, PIRLS 2016 reported that fourth-grade girls had a higher average reading score than boys (Mullis et al., 2017). This performance discrepancy could partly be a byproduct of gender difference in omitting behavior. As discussed earlier, most educational testing programs, including the PIRLS, treat the omitted response as an incorrect answer (Martin et al., 2016, 2017). As a result, boys, who tended to omit the items more frequently, will have lower scores by design. Thus, boys' abilities could be systematically underestimated and reported as underperforming.

This study also showed that test language affected nonresponse behaviors. Students who took the French version of the test tended to exhibit more not-reached and omitted responses. These differences may be closely related to translation issues. In international large-scale educational tests, tests are translated into various languages to be administered internationally. However, it has been reported that translation may cause confusion due to factors such as cultural inappropriateness, unclear interpretations of translated items, improper language uses, and item writing conventions (Grisay, 2002; Solano-Flores et al., 2009; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003). Reading tests like the PIRLS, in particular, can be more susceptible to these translation confusions, resulting in more frequent item nonresponse behaviors.

The findings on the item location suggested that students tended to not reach and omit the items that appeared later in the test, even after controlling for item difficulty and student ability. This finding is intuitively understandable because test-takers can be too exhausted to answer the questions or simply run out of time to complete the test, even though they are capable of

providing answers. Consequently, the knowledge/skills to be measured by the items located later in the test (typically more advanced and sophisticated ones) might not be properly evaluated and therefore underrepresented in the final test scores.

As for item format, we found that students tended to omit the constructed-response items more frequently than the multiple-choice items after controlling for item difficulty and student ability. This result was consistent with the findings by Okumura (2014). This difference may occur because the constructed-response items generally require more effort compared to the multiple-choice items. This could be a hassle for students who had low test motivation, in particular when the consequence is low-stake. Similar to the undesirable effect of item location, the knowledge/skills measured by the constructed-response items may not be properly assessed and hence underrepresented in the final test score.

The findings based on the PIRLS had several practical implications on test development. First, test developers should be cautious about the handling of not-reached and omitted responses. These item nonresponses are often treated as incorrect responses when computing students' proficiency scores (Martin et al., 2017). However, this method could systematically underestimate the ability of those students having frequent nonresponses by giving them lower test scores. To reduce such potential bias, researchers can consider some remedies in scoring, such as the imputation-based approach (e.g., Bernaards & Sijtsma, 2000; Huisman & Molenaar, 2001; Rubin, 1987; Sijtsma & van der Ark, 2003) and the model-based approach (Glas & Pimentel, 2008; Holman & Glas, 2005; Köhler et al., 2017; Pohl et al., 2014; Rose et al., 2010).

Second, item nonresponse behaviors can be alleviated by providing sufficient testing time. The current study found that students tended to not respond to questions closer to the end of the test. Moreover, students taking the French version were more likely to fail to complete the

test. This may suggest that not all students had sufficient testing time to answer all questions. For tests like the PIRLS, they are not designed to be a speed test (Martin et al., 2017). Hence, it is important to ensure the testing time is enough for students with different educational, language, and cultural backgrounds so that most, if not all, of the students, can respond to all the items.

Third, test developers could work on the design of constructed-response questions to mitigate the problem of frequent omitting. To give some examples, ensuring the prompts are concise and clear, requiring shorter rather than lengthy answers, and providing some structure to guide their response can be considered.

There are several directions for future studies. The explanatory IRTree model can be further extended to examine interaction effects between person and item properties (e.g., between gender and item format). This enables researchers to explore how person and item characteristics work synergistically to affect item nonresponse. Also, although the current study treated reading ability as a confounding factor, it is possible to inspect how ability interacts with the item/person property variables to affect the nonresponse. However, it should be noted that an unnecessarily complex model can lead to identification and convergence problems in estimation.

There are some caveats in interpreting the results of the current study. The findings of this study may not be generalized to other samples of the PIRLS. Gender, test language, item location, and item format may show different relationships to nonresponse behaviors in samples from other countries. To fully understand item nonresponse behaviors, it would be useful to conduct analysis across all countries as future research. Finally, we encourage further work to investigate whether similar findings will be found in other large-scale educational tests such as PISA, TIMSS, and NAEP.

Chapter 5: Discussion

This chapter will discuss the entire work presented in this dissertation. It will begin by summarizing this dissertation and the main findings. I will then highlight the contributions and novelties of this dissertation. Following that, I will discuss several issues that may arise in fitting the IRTree model, in particular, when using the *lme4* R package. Finally, I will point out the limitations of this dissertation and recommendations for future research.

5.1 Summary of this dissertation

This dissertation aimed to draw researchers' attention to the potential of the explanatory IRTree model in the study of response behaviors. In Chapter 2, I first introduced the IRTree model under the explanatory item response modeling framework and explicated the explanatory IRTree model within this framework. Based on that, in Chapter 3 and Chapter 4, I presented two studies, applying both the descriptive and explanatory IRTree models, for studying response styles and item nonresponse behaviors. As the main findings, Study-1 found the presence of two distinct extreme response styles and the acquiescence response style when responding to the Rosenberg's Self-esteem Scale. These findings had several critical implications on the score interpretation and use of this scale. Study-2 found several factors related to item nonresponse behaviors. Specifically, the findings showed that students taking the French version were more likely to leave the test incomplete and omit the items; boys were more likely to omit the items. Moreover, not-reached and omitted responses happened more frequently for the items closer to the end of the test. The omitted response was also more likely to happen to constructed-response items. These findings had practical implications for test development.

5.2 Contributions of this research

This dissertation makes new contributions in several aspects. First and foremost, this dissertation facilitates the explanatory use of the IRTree model. Despite the fact that the explanatory IRTree model is not entirely new (see Böckenholt, 2019; Debeer et al., 2017; Jeon & De Boeck, 2016; Okumura, 2014), it has been seldom discussed without a rigorous explanation. This dissertation is the first attempt to explicate the IRTree model within the explanatory item response modeling framework. By adopting this framework, I elaborated on how the standard descriptive IRTree model can be readily extended to the explanatory IRTree model as a form of GLNMM. This conceptualization makes it easier for applied researchers to understand and recognize the benefits of the explanatory IRTree model as long as they have some basic training in mixed models.

Moreover, I illustrated the explanatory IRTree model with real data examples and R codes in the context of studying response styles and item nonresponse behaviors. This didactic work not only can help applied researchers understand the explanatory IRTree model but also provides useful hands-on aids. In practice, this can help expand the versatility of the IRTree model and gain an in-depth understanding of response behaviors.

The two empirical studies in this dissertation have substantive contributions to the field of measurement and test development as well. Although the IRTree model has been used to inspect response styles, the previous applications often focused on the extreme and mid-point response styles only (Böckenholt, 2017; Böckenholt & Meiser, 2017; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013; Zettler et al., 2016). This study is the first application of the IRTree model to examine acquiescence and disacquiescence response styles based on item keying property (positively and negatively keyed items).

Moreover, this study revealed the existence of two distinct extreme response styles and examined them at both the scale and item levels. These findings deepen our understanding of response styles and caution the validity of the Rosenberg's Self-esteem Scale.

Likewise, Study-2 has several methodological and practical contributions. Although Okumura (2014) was the first to use the IRTree model to study person and item characteristics associated with item nonresponse, it was limited to omitted response only. This study is the first attempt to examine the effects of person and item characteristics on *both* not-reached and omitted responses simultaneously using an explanatory IRTree model. In addition, this study contributes to the measurement literature in terms of understanding item nonresponse. Using the PIRLS data, this study found extra evidence for the effects of gender and item format on item nonresponse. Moreover, this study identified two new factors of test language and item location. Based on these findings, I also provided suggestions for handling these undesirable effects such as treatment of item nonresponse, testing time, and the writing of constructed-response items.

5.3 Fitting the IRTree model using the *lme4* R package

In this dissertation, I fitted the IRTree model using the *lme4* package. The *lme4* package was developed for fitting a generalized linear mixed model and it has been widely used (Bates et al., 2015). Previous studies demonstrated how various item response models including the IRTree model can be estimated using the *lme4* functions (De Boeck et al., 2011; De Boeck & Partchev, 2012; Doran et al., 2007). The *lme4* R codes formulated in this dissertation were largely driven by this previous research. In the following, I will address several issues that may arise in using this package.

Readers may notice that the intercept is suppressed in the specification of the IRTree model in the *lme4*. That is, the R code includes 0 (or -1) along with predictors of interest to suppress the intercept. In item response models, this specification is commonly adopted to avoid one of the predictor values to be taken as the reference point (e.g., De Boeck et al., 2011; Doran et al., 2007), which may lead to cumbersome interpretation. For example, with the intercept, the fixed effect of the item will be an estimate of how much each item deviates from the reference item (e.g., first item). By suppressing the intercept, the fixed effect can be estimated for all the items without comparing them to a reference item. For this reason, the intercept was suppressed in the IRTree model.

Among applied researchers, one of the common issues with using *lme4* is that it does not always provide a p-value. However, this issue does not apply when fitting the IRTree model. The IRTree model was fitted using the *glmer* function of *lme4* as a "generalized" model. In this function, the p-value is calculated based on asymptotic Wald tests following the standard practice for generalized linear models (Bates et al., 2020). Even so, the developers pointed out that "these tests assume that shape of the log-likelihood surface and the accuracy of a chi-squared approximation to differences in log-likelihoods." Hence, researchers should be cautious about interpreting the p-value when these assumptions fail.

Finally, it should be noted that researchers can consider other software for fitting the IRTree model. Although this dissertation employed the *lme4* R package, any software with GLNMM available can be used to fit the IRTree model. Specifically, the *flirt* R package developed by Jeon and Rijmen (2016) can be a good option for applied researchers. This package is specialized for the IRTree model and includes functions that help researchers with data preparation and model specification.

5.4 Limitations and recommendations for future research

There are some caveats and limitations in this dissertation. First, the IRTree models demonstrated in this dissertation were limited to one-parameter models with binary decision outcomes. It is important to note that it is also possible to fit two- and three-parameter models (Jeon & De Boeck, 2016). In such cases, the IRTree model will be a generalized *nonlinear* mixed model. Equally worth noting, the nodes do not have to be binary (two branches). A node can have multiple branches depending on the researchers' questions and the data at hand. In these cases, multinomial or ordinal IRTree models are most appropriate, and the statistical packages that can handle multiple node outcomes such as the *flirt* R package by Jeon and Rijmen (2016) can be considered. In future research, it would be valuable to extend the explanatory IRTree model with a nonlinear form and multiple decision outcomes. These extensions will provide even more flexibility in tailoring the IRTree model to the researcher's specific data for their specific questions.

Second, the legitimacy of the assumed decision process in the tree structure is pivotal for the trustworthiness of the findings. If the decision process assumed by the researchers does not reflect the reality of how the respondents come to the response outcomes, the findings could be called into question. Hence a well-thought-out process is crucial to IRTree model applications. For future research, I recommend the following methods for choosing a decision process. First, researchers can hypothesize several alternative decision processes and evaluate their appropriateness based on goodness-of-fit measures such as – 2 log likelihood, AIC, and BIC. This approach helps identify the best decision process for the data. Recently, Debeer et al. (2017) used this approach and identified the best fitting process for studying item nonresponse. Second, researchers can evaluate the decision process by model residuals. Once the IRTree model is

fitted based on a postulated decision process, researchers can obtain residuals (i.e., the difference between the observed data and estimated data). The residuals can be useful for examining whether the hypothesized decision process goes along with the real data. Large residuals for many respondents suggest that the decision process is probably not suitable for the sample. Lastly, qualitative approaches, such as the think-aloud method, are commonly used to find out the realistic response processes (e.g., Charters, 2010; Fonteyn et al., 1993). Researchers can work with some of the participants using this approach and compare whether the results are consistent with the specified decision process. Qualitative approaches would be also beneficial when it comes to identify an appropriate decision process when researchers do not have a firm hypothesis. This approach can identify more than one process among respondents. In such a case, researchers can test all the different hypotheses against the data to see which one fits better to the sample data. Alternatively, one can fit separate IRTree models to reflect the population heterogeneity in the item response decision processes (e.g., Kim & Bolt, 2021).

In future research, I also recommend extending Study-1 and Study-2 for other different contexts. In Study-1, I presented the application of the IRTree model for studying response styles in the Likert-type Rosenberg's Self-esteem Scale. Future research can extend this application to other Likert-type psychological scales and examine response styles. Likewise, the application introduced in Study-2 can be used in other large-scale educational tests, such as PISA, TIMSS, and NAEP, to inspect item nonresponse. The method in Study-2 will be also useful for examining nonresponse behaviors in psychological measures or surveys. In these types of measures, it is common to include response options such as 'Do not know', 'Not applicable', and 'Reject to answer'. These response options can be readily modeled via a tree structure in the IRTree model. For example, researchers can first specify the decision query about whether the

question is applicable to the respondents (e.g., the question 'Are you pregnant?' for males). If the question is applicable, it can further specify whether the respondents provide a valid response or choose 'Don't know' or 'Reject to answer' as a second step. By modeling the response options in this way, it can help understand the underlying mechanism in item nonresponse.

5.5 Conclusion

This dissertation sheds light on the explanatory IRTree model and its applications in the context of studying response behaviors. I believe that this work will help researchers gain a deeper understanding of the IRTree model and learn more about response styles and item nonresponse behaviors. Measurement professionals and psychometricians, in particular, can benefit from learning this technique, for the purposes of test evaluation, validation, and research. Given its flexibility and versatility, researchers interested in studying response behaviors are also encouraged to utilize the IRTree model as they see fit.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Austin, J. S. (1992). The detection of fake good and fake bad on the MMPI-2. *Educational and Psychological Measurement*, *52*(3), 669–674.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *The Public Opinion Quarterly*, *48*(2), 491–509. https://doi.org/10.1086/268845

Bagley, C., & Mallick, K. (2001). Normative data and mental health construct validity for the Rosenberg Self-Esteem Scale in British adolescents. *International Journal of Adolescence and Youth*, *9*(2), 117–126. https://doi.org/10.1080/02673843.2001.9747871

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology*, *87*(2), 266–281. https://doi.org/10.1037/0022-3514.87.2.266

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2020). *Linear mixed-effects models using "Eigen" and S4*. https://github.com/lme4/lme4/

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

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*(3), 321–364. https://doi.org/10.1207/S15327906MBR3503_03

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

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

Böckenholt, U. (2019). Assessing item-feature effects with item response tree models. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 486–500. https://doi.org/10.1111/bmsp.12163

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/10.1111/bmsp.12086

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*(5), 335–352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814–833. https://doi.org/10.1177/0013164410388411

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118.

Charters, E. (2010). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education*, *12*(2). https://doaj.org/article/abaaa51b144d4a0c8c47ac83f290e99a

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading*, *10*(4), 331–362. https://doi.org/10.1207/s1532799xssr1004_1

Cho, S., Brown-Schmidt, S., De Boeck, P., & Shen, J. (2020). Modeling intensive polytomous time-series eye-tracking data: A dynamic tree-based item response model. *Psychometrika*, *85*(1), 154–184. https://doi.org/10.1007/s11336-020-09694-6

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (Eds.). (2003). Outliers and multicollinearity: Diagnosing and solving regression problems II. In *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed, pp. 390–430). L. Erlbaum Associates.

Coleman, C. (2013). Effects of negative keying and wording in attitude measures: A mixed-methods study (Doctoral dissertation). https://commons.lib.jmu.edu/diss201019/73/?utm_source=commons.lib.jmu.edu%2Fdiss201019%2F73&utm_medium=PDF&utm_campaign=PDFCoverPages

Cresswell, J., Schwantner, U., & Waters, C. (2015). *A review of international large-scale assessments in education: Assessing component skills and collecting contextual data*. Organization for Economic Cooperation and Development. https://doi.org/10.1787/9789264248373-en

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. https://doi.org/10.1177/001316444600600405

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28. https://doi.org/10.18637/jss.v039.i12

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 Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

De Leeuw, E. D., Hox, J. J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *19*(2), 153–176.

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, *54*(3), 333–363. https://doi.org/10.1111/jedm.12147

Di Chiacchio, C., De Stasio, S., & Fiorilli, C. (2016). Examining how motivation toward science contributes to omitting behaviours in the Italian PISA 2006 sample. *Learning and Individual Differences*, *50*, 56–63. https://doi.org/10.1016/j.lindif.2016.06.025

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/10.1016/j.intell.2016.02.012

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 Package. *Journal of Statistical Software*, *20*(2), 1–28. https://doi.org/10.18637/jss.v020.i02

Finlay, W. M., & Lyons, E. (2001). Methodological issues in interviewing and using self-report questionnaires with people with mental retardation. *Psychological Assessment*, *13*(3), 319–335. https://doi.org/10.1037//1040-3590.13.3.319

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research*, *3*(4), 430–441. https://doi.org/10.1177/104973239300300403

Foulds, M. L., & Warehime, R. G. (1971). Effects of a" fake good" response set on a measure of self-actualization. *Journal of Counseling Psychology*, *18*(3), 279–280.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907–922. https://doi.org/10.1177/0013164408315262

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

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

Grisay, A. (2002). Translation and cultural appropriateness of the test and survey material. In R. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 57–70). Organization for Economic Cooperation and Development.

Groves, R. M. (1989). Nonresponse in sample surveys. In *Survey Errors and Survey Costs* (pp. 133–183). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471725277.ch4

Hagborg, W. J. (1993). The Rosenberg Self-Esteem Scale and Harter's Self-Perception Profile for adolescents: A concurrent validity study. *Psychology in the Schools*, *30*(2), 132–136.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education*, *43*(1), 40–46.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*(4), 665–686. https://doi.org/10.1177/0013164411430707

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Huisman, M. (1999). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. DSWO Press.

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). Springer. https://doi.org/10.1007/978-1-4613-0169-1_13

Hurley, J. R. (1998). Timidity as a response style to psychological questionnaires. *The Journal of Psychology*, *132*(2), 201–210. https://doi.org/10.1080/00223989809599159

Hyland, P., Boduszek, D., Dhingra, K., Shevlin, M., & Egan, A. (2014). A bifactor approach to modelling the Rosenberg Self Esteem Scale. *Personality and Individual Differences*, *66*, 188–192. https://doi.org/10.1016/j.paid.2014.03.034

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*(4), 243–252. https://doi.org/10.1037/h0045996

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

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

Jeon, M., & Rijmen, F. (2016). A modular approach for item response theory modeling with the R package flirt. *Behavior Research Methods*, *48*(2), 742–755. https://doi.org/10.3758/s13428-015-0606-z

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

Kam, C. C. S., & Fan, X. (2017). Approaches to handling common response styles and issues in educational surveys. *Oxford Research Encyclopedia of Education*. https://doi.org/10.1093/acrefore/9780190264093.013.90

Khorramdel, L., & Davier, M. von. (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

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

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychology Science*, *57*(4), 499–522.

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397–419. https://doi.org/10.1111/jedm.12154

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (Report No. CSE Technical Report 357). Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://eric.ed.gov/?id=ED378220

Laffitte, R. G. (1984). Effects of item order on achievement test scores and students' perception of test difficulty. *Teaching of Psychology*, *11*(4), 212–214. https://doi.org/10.1177/009862838401100405

LaHuis, D. M., Blackmore, C. E., Bryant-Lees, K. B., & Delgado, K. (2019). Applying item response trees to personality data in the selection context. *Organizational Research Methods*, *22*(4), 1007–1018. https://doi.org/10.1177/1094428118780310

Lavrakas, P. (2008a). Item order randomization. In *Encyclopedia of survey research methods* (Vols. 1–0). Sage Publications, Inc. https://doi.org/10.4135/9781412963947.n255

Lavrakas, P. (2008b). Self-reported measure. In *Encyclopedia of survey research methods* (Vols. 1–0). Sage Publications, Inc. https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods

Logan, S., & Johnston, R. (2010). Investigating gender differences in reading. *Educational Review*, *62*(2), 175–187. https://doi.org/10.1080/00131911003637006

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Matters, G., & Burnett, P. C. (1999). Multiple-choice versus short-response items: Differences in omit behaviour. *Australian Journal of Education*, *43*(2), 117–128. https://doi.org/10.1177/000494419904300202

Matters, G., & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement*, *63*(2), 239–256. https://doi.org/10.1177/0013164402250988

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

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/10.1016/j.paid.2008.01.010

Min, H., Zickar, M., & Yankov, G. (2018). Understanding item parameters in personality scales: An explanatory item response modeling approach. *Personality and Individual Differences*, *128*, 1–6. https://doi.org/10.1016/j.paid.2018.02.012

Mislevy, R. J., & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Report No. RR-96-30-ONR). Educational Testing Service. http://doi.wiley.com/10.1002/j.2333-8504.1996.tb01708.x

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

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity*, *42*(6), 779–794. https://doi.org/10.1007/s11135-006-9067-x

Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research*, *22*(1), 93–119. https://doi.org/10.1093/ijpor/edp036

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*(2), 271–298. https://doi.org/10.1080/1359432X.2010.550680

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools in 35 countries*. International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study primary schools in 40 countries*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Myers, A. J., Ames, A. J., Leventhal, B. C., & Holzman, M. A. (2020). Validating rubric scoring processes: An application of an item response tree model. *Applied Measurement in Education*, *33*(4), 293–308. https://doi.org/10.1080/08957347.2020.1789143

Okumura, T. (2014). Empirical differences in omission tendency and reading ability in PISA. *Educational and Psychological Measurement*, *74*(4), 611–626. https://doi.org/10.1177/0013164413516976

Parker, J. F., & Ryan, V. (1993). An attempt to reduce guessing behavior in children's and adults' eyewitness identifications. *Law and Human Behavior*, *17*(1), 11–26.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32. 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 social psychological attitudes* (pp. 17–59). Academic Press. 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

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*(3), 423–452. https://doi.org/10.1177/0013164413504926

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325.

Randall, J., Cheong, Y. F., & Engelhard Jr, G. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement*, *71*(1), 129–147.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Reynolds, N., & Smith, A. (2010). Assessing the impact of response styles on cross-cultural service quality evaluation: A simplified approach to eliminating the problem. *Journal of Service Research*, *13*(2), 230–243. https://doi.org/10.1177/1094670509360408

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185–205. https://doi.org/10.1037/1082-989X.8.2.18

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Doctoral dissertation). https://www.db-thueringen.de/receive/dbt_mods_00022476

Rose, N., Davier, M. von, & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Report No. ETS RR-10-11). Educational Testing Service. http://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.2010.tb02218.x

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, 34, 1-97. https://doi.org/10.1007/BF03372160

Schneider, S. (2016). Extracting response style bias from measures of positive and negative affect in aging research. *Journals of Gerontology: Series B*, *73*(1), 67–74. https://doi.org/10.1093/geronb/gbw103

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*(4), 505–528. https://doi.org/10.1207/s15327906mbr3804_4

Silverstein, M., & Bengtson, V. L. (2008). *Longitudinal Study of Generations, California, 1971, 1985, 1988, 1991, 1994, 1997, 2000, 2005*. Inter-university Consortium for Political and Social Research [distributor].

Sinclair, S. J., Blais, M. A., Gansler, D. A., Sandberg, E., Bistis, K., & LoCicero, A. (2010). Psychometric properties of the Rosenberg Self-Esteem Scale: Overall and across demographic groups living within the United States. *Evaluation & the Health Professions*, *33*(1), 56–80. https://doi.org/10.1177/0163278709356187

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing*, *9*(2), 78–91. https://doi.org/10.1080/15305050902880835

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573. https://doi.org/10.1002/tea.1018

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3–13.

Stanke, L., & Bulut, O. (2019). Explanatory item response models for polytomous item responses. *International Journal of Assessment Tools in Education*, *6*(2), 259–278. https://doi.org/10.21449/ijate.515085

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522–547. https://doi.org/10.3102/1076998613481500

UNESCO Institute for Statistics. (2012). *International standard classification of education: ISCED 2011*. UNESCO Institute for Statistics.

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, *47*(1), 157–172. https://doi.org/10.1509/jmkr.47.1.157

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

Vermillion, M., & Dodder, R. A. (2007). An examination of the Rosenberg Self-Esteem Scale using collegiate wheelchair basketball student athletes. *Perceptual and Motor Skills*, *104*(2), 416–418. https://doi.org/10.2466/pms.104.2.416-418

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross - Cultural Psychology*, *34*(6), 702. https://doi.org/10.1177/0022022103257070

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, *33*(5), 352–364. https://doi.org/10.1027/1015-5759/a000291

Wu, Y., & Jin, K. (2020). An extended item response tree model for wording effects in mixed-format scales (paper presentation). In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J. Kim (Eds.), *Quantitative Psychology. IMPS 2019* (Vol. 322). Springer. https://doi.org/10.1007/978-3-030-43469-4_4

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports. *Journal of Personality*, *84*(4). https://doi.org/10.1111/jopy.12172

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*(3), 10–16. https://doi.org/10.1111/j.1745-3992.1991.tb00198.x

Appendices

Appendix A

Rosenberg's Self-esteem Scale

1. WB033R8 – I feel that I'm a person of worth, at least on an equal basis with others

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

2. WB034R8 – I feel that I have a number of good qualities

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

3. WB035R8 – All in all, I am inclined to feel that I am a failure

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

4. WB036R8 – I am able to do things as well as most other people

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

5. WB037R8 – I feel I do not have much to be proud of

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

6. WB038R8 - I take a positive attitude toward myself

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

7. WB039R8 – On the whole, I am satisfied with myself

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

8. WB040R8 – I wish I could have more respect for myself

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

9. WB041R8 – I certainly feel useless at times

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

10. WB042R8 – At times I think I am no good at all

□ Strongly agree □ Agree □ Disagree □ Strongly disagree

Appendix B

Table B.1 Long format of an example data matrix for fitting the IRTree model in Chapter 4

Person	Item (original responses)	Node	Node outcome
1	Item-1 (Correct)	Node-1	0
1	Item-1 (Correct)	Node-2	0
1	Item-1 (Correct)	Node-3	1
1	Item-2 (Incorrect)	Node-1	0
1	Item-2 (Incorrect)	Node-2	0
1	Item-2 (Incorrect)	Node-3	0
1	Item-3 (Not-reached)	Node-1	1
1	Item-3 (Not-reached)	Node-2	0
1	Item-3 (Not-reached)	Node-3	0
1	Item-4 (Not-reached)	Node-1	NA
1	Item-4 (Not-reached)	Node-2	NA
1	Item-4 (Not-reached)	Node-3	NA
1	Item-5 (Not-reached)	Node-1	NA
1	Item-5 (Not-reached)	Node-2	NA
1	Item-5 (Not-reached)	Node-3	NA
1	…		
1	Item-10 (D)	Node-1	NA
1	Item-10 (D)	Node-2	NA
1	Item-10 (D)	Node-3	NA

2	Item-1 (Incorrect)	Node-1	0
2	Item-1 (Incorrect)	Node-2	0
2	Item-1 (Incorrect)	Node-3	0
2	Item-2 (Correct)	Node-1	0
2	Item-2 (Correct)	Node-2	0
2	Item-2 (Correct)	Node-3	1
	…		

Note. This data format is based on the mapping matrix in Table 4.1; NA = not applicable. For each person, if the first not-reached response was identified in the test, the rest of the items were treated as NA. This is because the test-taker stops attempting the test and the remaining items are not applicable anymore.

Appendix C

This appendix presents the results of the explanatory IRTree model with the second random half in Chapter 4.

As shown in Table C.1, the explanatory IRTree model fitted better than the descriptive model. The conditional means of the three nodes were estimated ($\beta_{Node-1} = -6.443$, $\beta_{Node-2} = -4.42$, and $\beta_{Node-3} = 0.797$). The random effects were student's ability and item difficulty ($\theta_{Node-3} = 1.151$ and $u_{Node-3} = 2.030$) for Node-3 and additional item variabilities ($u_{Node-1} = 1.077$ and $u_{Node-2} = 0.527$) for Node-1 and Node-2. The effects of person and item characteristics were reported in Table C.2.

Table C.1 Model fits of descriptive and explanatory IRTree models (with the second random half)

Models	-2 LL	AIC	BIC
1. Descriptive IRTree model	353942.6	353962.6	354078.4
2. Explanatory IRTree model	352801.9	352835.9	353032.8

Note. -2 LL = -2 times log likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion. A lower value of -2 LL, AIC, and BIC indicates that the model fits better to the data.

Table C.2 The fixed effects of the person and item characteristics in the explanatory IRTree model (with the second random half)

Node-1 (not-reached)		β	*SE*	*p*	*Odds ratio*
	Item location	1.190	0.090	<.001	3.29
	Gender	0.037	0.049	0.450	1.04
	Test language	0.530	0.050	<.001	1.70
Node-2 (omitted)		β	*SE*	*p*	
	Item location	0.239	0.045	<.001	1.27
	Item format	0.972	0.089	<.001	2.64
	Gender	0.097	0.021	<.001	1.10
	Test language	0.542	0.021	<.001	1.72

Appendix D

The IRTree models were fitted using a glmer function in the *lme4* R package. Corresponding R codes for the IRTree models in Chapter 3 and Chapter 4 were specified as follows. Note that DV is the node outcomes. Item denotes item indicator variable and person denotes person indicator variable.

lme4 R codes in Chapter 3

Descriptive IRTree model

glmer(DV~ 0 + nodes:item + (0 + nodes|person))

Explanatory IRTree model

glmer(DV~ 0 + nodes:item_keying + (0 + nodes|person) + (0+nodes|item)).

lme4 R codes in Chapter 4

Descriptive IRTree model

glmer(DV ~ 0 + node1 + node 2 + node3 + (0 + node3|person) + (0 + node1 + node2 + node3|item)).

Explanatory IRTree model

glmer(DV ~ 0 + node1 + node 2 + node3 + node1:(item_ location + gender + test_language) + node2:(item_location + gender + test_language + item_format) + (0 + node3|person) + (0 + node1 + node2 + node3|item)).