UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Extensions and Applications of Item Explanatory Models to Polytomous Data in Item Response Theory

Permalink
https://escholarship.org/uc/item/4pf2p2fj

Author
Kim, Jinho

Publication Date
2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Extensions and Applications of Item Explanatory Models to Polytomous Data in Item Response Theory

by

Jinho Kim

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Sophia Rabe-Hesketh
Professor Alan Hubbard

Summer 2018

Extensions and Applications of Item Explanatory Models to Polytomous Data in Item Response Theory

Copyright 2018

by

Jinho Kim

Abstract

Extensions and Applications of Item Explanatory Models to Polytomous Data in Item Response Theory

by

Jinho Kim

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

Polytomous item explanatory models aim to provide informative feedback for improving item design and generation, test development, and educational assessment so that they can help improve teachers' instruction and students' learning ultimately. In order to demonstrate methodological advantages and practical implications of these models, this dissertation addresses extensions and applications of item explanatory models to polytomous data, using the Linear Logistic Test Model (LLTM; Fischer, 1973) approach, in the context of explanatory measurement (De Boeck & Wilson, 2004). The three papers in this dissertation are intended to be separately publishable, but they share the common theme of explanatory and predictive inferences from the ordered-category item responses.

The first paper in Chapter 2 discusses the applicability of the Many-Facet Rasch Model (MFRM; Linacre, 1989) to polytomous item explanatory models, considering categorical item properties as facets or sub-facets within the item facet. To demonstrate practical applications of the MFRM-based item explanatory model, two empirical studies investigate how item properties explain and predict the overall item difficulties in the Carbon Cycle assessment data and in the Social Evaluative Reasoning in the workplace data respectively. The results suggest that the MFRM-based item explanatory models enable to check and validate underlying hypotheses applied to the item design and construction as well as to develop a new item in a scientific and systematic manner rather than an intuitive manner. The results serve as helpful feedback for enhancing quality of the assessment and instrument development.

The second paper in Chapter 3 investigates polytomous item explanatory models based on the adjacent-categories logits under the multivariate generalized linear mixed modeling framework. Building on the original ideas of the MFRM and the Linear Partial Credit Model (LPCM; Fischer & Ponocny, 1994), polytomous Rasch family models are extended to two item explanatory versions using the MFRM and the LPCM approaches. To demonstrate the practical differences between the two polytomous item explanatory approaches, two empirical studies in this paper examine how item properties explain and predict the overall item difficulties or the step difficulties

each in the Carbon Cycle assessment data and in the verbal aggression data. The results suggest that the two polytomous item explanatory models are methodologically and practically different in terms of the target difficulty parameters of polytomous items which are explained by item properties, the types of item properties incorporated in the design matrix, and the types of item property effects.

The third paper in Chapter 4 proposes three polytomous item explanatory models with random item errors, based on the models specified in the second paper. As in the Linear Logistic Test Model with item error (LLTM + ε; Mislevy, 1988; Janssen, Schepers, & Peres, 2004) approach, the proposed polytomous random item effects models can take the uncertainty in explanation and/or the random nature of item parameters into account for polytomous items. For estimation of the proposed models with crossed random effects, available estimation methods are reviewed, and a Bayesian inference is adopted. To examine how the proposed models function in different explanatory conditions, two simulation studies evaluate model comparison and parameter recovery as well as the effects of model misspecification. In addition, two empirical studies demonstrate practical implications and applications of the proposed models to two real data sets, the Carbon Cycle assessment data and the verbal aggression data. The simulation findings suggest that the proposed polytomous item explanatory models with random item errors perform better than the models without item error terms for making accurate statistical inferences, such as estimation or hypothesis testing for the item property effects. The empirical findings indicate that the proposed models outperform the models without random item errors in terms of the goodness-of-fit and reconstructing the step difficulties, and also show that methodological and practical differences between the two polytomous item explanatory approaches.

Keywords: explanatory measurement, item property, item explanatory model, polytomous data, linear logistic test model, many-facet Rasch model, linear partial credit model, linear logistic test model with item error, random item effects, random item error

Dedication

To those who encouraged, helped, and fed me along the way.

Acknowledgements

First and foremost, I would like to deeply thank my advisor, Professor Mark Wilson. This work would not have been possible without his unwavering support, intellectual advice, careful and keen comments, and encouragement. I would also like to thank my second advisor, Professor Sophia Rabe-Hesketh. Her warm and encouraging support, sharp comments, and brilliant suggestions made this work completed. Also, I would like to thank Professor Alan Hubbard for his outstanding instruction and encouragement and for providing an outside perspective.

My gratitude is also due to Doctor Karen Draney, for her support, guidance, and counsel while working with her at the BEAR center. I would also like to thank Professor David Pearson for his serving on a chair for my qualifying exam committee and for his encouraging support. My gratitude also extends to my former advisor, Professor Sehee Hong, for initiating my exciting academic journey, and to professors in the Department of Education, Korea University.

I am deeply indebted to all my dear friends in the QME/SRM program, in GSE, in UC Berkeley, and in every community I belong to as well as in Korea. A page is just not enough to name them all, but I really thank each and every one of them here. Without their warm support and help, I would never have completed my doctoral course over six years. Though they remain nameless, they know who they are.

Lastly but most importantly, I sincerely appreciate all my family and relatives for their endless love and unconditional support throughout my life. This dissertation is especially dedicated to my parents, Jaeik Kim and Minja Suh, and to my brother's family as well as to my wife's family. Most of all, nonetheless, the deepest appreciation must go to my beloved wife, Yumi Park, and to our most adorable baby waiting to come into the world. Yumi, without your greatest love and support, I would have got nowhere. Also, to my soon-to-be-born daughter, "Congratulations", see you soon. Your mom and I are delighted to give you your real name very soon.

Table of Contents

Dedication ..... **i**

Acknowledgements ..... **ii**

Table of Contents ..... **iii**

List of Tables ..... **vi**

List of Figures ..... **viii**

Chapter 1. Overall Introduction **1**

Chapter 2. Investigation of Item Property Effects on Polytomous Items Using Item Explanatory IRT Models **4**

2.1. Introduction 4

2.2. Item Explanatory IRT Models under the Multidimensional Random Coefficient Multinomial Logit Model Framework 6

2.2.1. The Multidimensional Random Coefficient Multinomial Logit Model Framework 6

2.2.2. Item Response Models for Polytomous data 8

2.2.3. Item Explanatory Models Using the Many-Facet Rasch Model 9

2.3. Empirical Studies 11

2.3.1. Empirical Data Analysis 11

2.3.2. Application to the Carbon Cycle Assessment data 13

2.3.3. Application to the Social Evaluative Reasoning data 22

2.4. Conclusion and Discussion 31

Chapter 3. Polytomous Item Explanatory IRT Models **34**

3.1. Introduction 34

3.2. Polytomous Extensions of the Rasch Model under the Multivariate Generalized Linear Mixed Modeling Framework 36

3.2.1. The Multivariate Generalized Linear Mixed Modeling Framework 36

3.2.2. Polytomous Rasch Models 37

3.3. Item Explanatory Extensions of Polytomous Rasch Models under the Multivariate Generalized Linear Mixed Modeling Framework 41

3.3.1. Item Explanatory Extensions with Item Properties 41

3.3.2. Many-Facet Rasch Model (MFRM) with a Decomposition of the Item Location Parameters 43

3.3.3. Linear Partial Credit Model (LPCM) with a Decomposition of the Step Difficulty Parameters 45
3.4. Empirical Studies 48
3.4.1. Empirical Data Analysis 48
3.4.2. Application to the Carbon Cycle Assessment data 49
3.4.3. Application to the Verbal Aggression data 58
3.5. Conclusion and Discussion 66
Chapter 4. Polytomous Item Explanatory IRT Models with Random Item Effects 68
4.1. Introduction 68
4.2. Polytomous Item Response Models with Random Item Effects 70
4.2.1. Rasch Family Models for Polytomous data 70
4.2.2. Polytomous Rasch Models with Random Item Effects 71
4.3. Polytomous Item Explanatory Models with Random Item Errors 75
4.3.1. Polytomous Item Explanatory Models 75
4.3.2. Polytomous Item Explanatory Models with Random Item Errors 77
4.4. Estimation of Polytomous Random Item Effects Models 80
4.4.1. Estimation of Crossed Random Effects Models 80
4.4.2. Bayesian methods for Polytomous Item Explanatory Models with Random Item Errors 82
4.5. Simulation Studies 85
4.5.1. Simulation Design and Data Generation 85
4.5.2. Study 1: Model Comparison 88
4.5.3. Study 2: Parameter Recovery 99
4.6. Empirical Studies 108
4.6.1. Application to the Carbon Cycle Assessment data 109
4.6.2. Application to the Verbal Aggression data 116
4.7. Conclusion and Discussion 122
References 125
Appendices 132
A.1. Linear Rating Scale Model (LRSM) with a Decomposition of the Item Location Parameters for Chapter 3 132

A.2. *gllamm* Syntax for Polytomous Item Explanatory IRT Models for Chapter 3 134
A.3. *Stan* Syntax for Polytomous Item Explanatory Models with Random Item Errors for Chapter 4 136

List of Tables

Table 2-1. The Learning Progression Framework for the Carbon Cycle items 13
Table 2-2. Q matrix for the Carbon Cycle items composed of three item properties 15
Table 2-3. Goodness-of-fit of item explanatory models to the Carbon Cycle assessment data ... 16
Table 2-4. Step difficulties for the Carbon Cycle items in item explanatory models 17
Table 2-5. Item property effects on the Carbon Cycle items and item fit in the interaction MFRM 19
Table 2-6. Q matrix for the evaluative inferencing items composed of three item properties 25
Table 2-7. Goodness-of-fit of item explanatory models to the Social Evaluative Reasoning data 26
Table 2-8. Step difficulties for the evaluative inferencing items in item explanatory models 27
Table 2-9. Item property effects on the evaluative inferencing items and item fit in the additive MFRM 29
Table 3-1. The Learning Progression Framework for the Carbon Cycle assessment 49
Table 3-2. The Carbon Cycle items composed of three item properties 50
Table 3-3. Performance of the MFRM and the LPCM on the Carbon Cycle assessment data 53
Table 3-4. Correlations of the estimated and calculated step difficulties in the Carbon Cycle items 55
Table 3-5. Item property effects on the Carbon Cycle items in the MFRM variation 55
Table 3-6. Item property effects on the Carbon Cycle items in the LPCM variation 56
Table 3-7. The verbal aggression items composed of three item properties 59
Table 3-8. Performance of the MFRM and the LPCM on the verbal aggression data 61
Table 3-9. Correlations of the estimated and calculated step difficulties in the verbal aggression items 63
Table 3-10. Item property effects on the verbal aggression items in the MFRM variation 64
Table 3-11. Item property effects on the verbal aggression items in the LPCM variation 65
Table 4-1. Three types of polytomous random item effects for the PCM 75
Table 4-2. Three proposed polytomous item explanatory models with random item errors 78
Table 4-3. Parameter generating values for the MFRM + OIE 87
Table 4-4. Parameter generating values for the LPCM + SSE 88
Table 4-5. Goodness-of-fit of models fitted to the MFRM + OIE generation data 90
Table 4-6. Correlations of the estimated and calculated step difficulties between the models fitted to the MFRM + OIE generation data 91
Table 4-7. Goodness-of-fit of models fitted to the LPCM + SSE generation data 95
Table 4-8. Correlations of the estimated and calculated step difficulties between the models fitted to the LPCM + SSE generation data 96
Table 4-9. Parameter recovery for the MFRM + OIE-based simulation 101

Table 4-10. Item property effect parameters in the MFRM + OIE-based simulation 103
Table 4-11. Parameter recovery for the LPCM + SSE-based simulation 105
Table 4-12. Item property effect parameters in the LPCM + SSE-based simulation 108
Table 4-13. The Carbon Cycle items and three categorical item properties 109
Table 4-14. Goodness-of-fit of models fitted to the Carbon Cycle assessment data 110
Table 4-15. Correlations of the estimated and calculated step difficulties between the models fitted to the Carbon Cycle assessment data 111
Table 4-16. Model parameter estimates and standard errors in the MFRM-based polytomous item explanatory models to the Carbon Cycle assessment data 113
Table 4-17. Model parameter estimates and standard errors in the LPCM-based polytomous item explanatory models to the Carbon Cycle assessment data 115
Table 4-18. The verbal aggression items and three item design factors 117
Table 4-19. Goodness-of-fit of models fitted to the verbal aggression data 118
Table 4-20. Correlations of the estimated and calculated step difficulties between the models fitted to the verbal aggression data 118
Table 4-21. Model parameter estimates and standard errors in the MFRM-based polytomous item explanatory models to the verbal aggression data 120
Table 4-22. Model parameter estimates and standard errors in the LPCM-based polytomous item explanatory models to the verbal aggression data 121

List of Figures

Figure 2-1. Item explanatory models under the MRCML model framework 7
Figure 2-2. Flow chart of the scoring rubric for the item 1 (BODYTEMP) 14
Figure 2-3. Graphical comparison of step difficulties for the Carbon Cycle items in item explanatory models 18
Figure 2-4. Wright map for item property effects on the Carbon Cycle items 21
Figure 2-5. The 4th workplace scenario for understanding figurative language: sweet tooth idiom 23
Figure 2-6. Construct map for evaluative inferencing ability with corresponding outcome space 24
Figure 2-7. Graphical comparison of step difficulties for the evaluative inferencing items in item explanatory models 28
Figure 2-8. Wright map for item property effects on the evaluative inferencing items 30
Figure 3-1. Flow chart of the scoring rubric for the item 1 (BODYTEMP) 51
Figure 3-2. Graphical comparison of step difficulties between the models fitted to the Carbon Cycle assessment data 54
Figure 3-3. Graphical comparison of step difficulties between the models fitted to the verbal aggression data 63
Figure 4-1. Cross-classified item responses in a random item effects model 80
Figure 4-2. Goodness-of-fit of models fitted to the MFRM + OIE generation data 90
Figure 4-3. Graphical agreement between the models fitted to the MFRM + OIE generation data 92
Figure 4-4. Scatter plot with 95% credible intervals for the MFRM + OIE-based simulation 93
Figure 4-5. Goodness-of-fit of models fitted to the LPCM + SSE generation data 95
Figure 4-6. Graphical agreement between the models fitted to the LPCM + SSE generation data 97
Figure 4-7. Scatter plot with 95% credible intervals for the LPCM + SSE-based simulation 98
Figure 4-8. Parameter recovery for the MFRM + OIE-based simulation 102
Figure 4-9. Parameter recovery for the LPCM + SSE-based simulation 106
Figure 4-10. Graphical agreement between the models fitted to the Carbon Cycle assessment data 112
Figure 4-11. Graphical agreement between the models fitted to the verbal aggression data 119

Chapter 1.
Overall Introduction

Explanatory measurement

Advances in theories of learning, cognitive science, and human development as well as increased demands for educational assessments to reform and inform education lead us to consider the confluence of such theories and psychometrics, which encourage an integrative perspective on assessment practices for the design, administration, and use of the assessments (Pellegrino, Baxter & Glaser, 1999). To help students' academic success, assessment practices must be grounded in substantive theories of cognitive development and suitable standards to describe students' acquisition of proficiency and expertise in specific academic contents (Pellegrino et al., 1999). Educational assessments can provide substantial information to be used as feedback to improve teachers' instruction and students' learning (Black & Wiliam, 1998), and the strength of the feedback depends on the capability and quality of informative assessments.

In the item response theory (IRT) perspective, explanatory measurement (De Boeck & Wilson, 2004) can strengthen the feedback by enhancing explanatory inferences from the assessments. When we have item response data as the assessment outcomes, the person side and/or item side of the data can be investigated through psychometric models for explanatory measurement, namely explanatory item response models (EIRM; De Boeck & Wilson, 2004). By using person properties and/or item properties aligned with theoretical and practical factors of the assessments, EIRMs explain the effects of persons and/or items so that they can enrich explanations and inferences from the assessment outcomes.

Item explanatory models

Among person explanatory, item explanatory, and doubly explanatory models of the EIRM approach, this dissertation will mainly focus on item explanatory models in which item properties are incorporated to account for the item effects (De Boeck & Wilson, 2004). In particular, the Linear Logistic Test Model (LLTM; Fischer, 1973) is well known, widely used, and a general approach to item explanatory extensions (Kubinger, 2009; Rost, 2001; Rost & Langeheine, 1997). The LLTM approach, which decomposes the difficulties of specific items into a linear combination of elementary components is a matter of concern in this dissertation. This approach is useful for assessment practices to examine the effect of item properties such as testing conditions, item design variables, content-specific learning, and cognitive operations in various assessment contexts. Thus, item explanatory models can provide informative feedback for improving test development, item generation, and educational assessments.

In addition, it is often the case that the observed item properties cannot perfectly explain the item difficulties (De Boeck, 2008), even with substantive theories behind the item explanatory models. To avoid such unrealistic cases, it is appealing to add a random error term into the item regression component of the models. This approach is the Linear Logistic Test Model with item error (LLTM + ε; Mislevy, 1988; Janssen, Schepers, & Peres, 2004), which is a type of random

item IRT model (De Boeck, 2008). I will use both the LLTM approach and the LLTM + ε approach to extend item explanatory models to polytomous data. This work will be a methodological foundation for more extensive explanatory measurement to develop person explanatory or doubly explanatory models and their multidimensional extensions as well as explanatory two-parameter or three-parameter IRT models.

Polytomous item responses

In educational assessments, students' item responses are typically observed as categorical data, either dichotomous (binary) or polytomous (multi-categorical). Most of the application studies using item explanatory models have investigated dichotomous data rather than polytomous data (e.g. De Boeck, 2008; Fischer, 1973; Hartig, Frey, Nold, & Klieme, 2012; Kubinger, 2009; Poinstingl, 2009). Nevertheless, we encounter polytomous data frequently in a wide range of educational, psychological, and sociological applications (De Boeck & Wilson, 2004).

It is very common to have ordered-category responses as students' achievement outcomes in educational assessment contexts. For example, in educational assessment practices, a learning progression represents hypothetical descriptions of the successively more sophisticated levels of students' understanding and competency in their learning about a specific topic over time (Duncan & Hmelo-Silver, 2009). When an assessment is developed with the idea of a learning progression in mind, the levels of students' achievement within that learning progression are scored and provided in the form of ordered-category responses (e.g. Jin & Anderson, 2012; Jolin, 2015).

Given the ordered-category responses, adjacent-categories logit relationship will be employed for polytomous extensions of item explanatory models, referred to as polytomous item explanatory models in this dissertation. Most of ordered-category responses in educational assessment or cognitive development contexts are subjectively assigned scores between categories, and adjacent-categories logit ordinal regression models are suitable for those scores (Anderson, 1984; Johnson, 2007). Also, adjacent-categories logits are quite useful when contrasting probabilities of the responses in pairs of adjacent categories (Hartzel, Agresti, & Caffo, 2001), regarded as the local comparison of ordered categories (Masters & Wright, 1997).

The three papers

To address these concerns about explanatory measurement, item explanatory models, and polytomous item responses, I will investigate extensions and applications of item explanatory models to polytomous data in this dissertation, which is comprised of three related papers.

The first paper will address applications of an item explanatory IRT model to polytomous data. Using the Many-Facet Rasch Model (MFRM; Linacre, 1989) for item explanatory measurement, I will investigate how item properties can explain the overall item difficulties in the Carbon Cycle assessment data and in the Social Evaluative Reasoning in the workplace data as well.

The second paper will discuss polytomous item explanatory IRT models based on adjacent-categories logits under the multivariate generalized linear mixed modeling framework. Polytomous Rasch family models will be extended to their item explanatory versions, and then will be applied to the Carbon Cycle assessment data and the verbal aggression data empirically.

Lastly, the third paper will propose three polytomous item explanatory IRT models with random item errors, based on the models specified in the second paper. I will investigate how to impose random effects on the item parameters in polytomous Rasch family models, and then will add random item errors to the polytomous item explanatory IRT models. To demonstrate practical applications of the proposed models, I will conduct two simulation studies and two empirical studies on the Carbon Cycle assessment and the verbal aggression data sets, using a Bayesian estimation method.

Chapter 2.
Investigation of Item Property Effects on Polytomous Items Using Item Explanatory IRT Models

2.1. Introduction

In educational research, it is important that educational assessment and measurement can provide substantial information and feedback to enhance teachers' instruction and students' learning (Atkinson & Paulson, 1972; Black & Wiliam, 1998). In item response theory (IRT), the Rasch model is well known and widely used as an item response model which is effective at providing information about student achievements and item characteristics in terms of both quantitative analysis and qualitative interpretation. Furthermore, extensions of the Rasch model have been growing and are applied to more complicated measurement issues such as explanatory measurement, polytomous data, multidimensionality, and mixture distributions, as well as to practical uses for improving item design, construct validation, and examining rater effects (De Boeck & Wilson, 2004; Linacre, 1989; Rost, 2001). The Rasch family models which are extended for explanatory measurement and polytomous data are the focus in this paper. These extended Rasch family models enhance explanatory inferences from various item responses so that they can enrich informative feedback for students and teachers as well as for educational researchers.

Among the extended Rasch family models (Rost, 2001), the Linear Logistic Test Model (LLTM; Fischer, 1973) is a typical approach to item explanatory measurement, regarded as an item explanatory IRT model. It was developed to help explain and predict the likelihood of correctly answering a particular item by incorporating item properties, such as item content features. This approach decomposes item difficulty parameters of the Rasch model into linear combinations of elementary components related to the item properties or features (Embretson & Reise, 2000). The LLTM is a valuable tool when generating items, because it can test a hypothesis about which item properties make an item more difficult, allowing for further investigation of construct validity (Embretson & Gorin, 2001; Poinstingl, 2009). When a specific rationale is established for creating items in a certain way, the LLTM can assess the appropriateness of that rationale, helping enable automatized item generation (Holling, Bertling, & Zeuch, 2009; Reif, 2012). The LLTM approach can also be used effectively for psychometric studies on measuring the effect of various testing conditions such as item presentation position, content-specific learning, speeded item presentation, and item response format, as well as measurement of change (Cristante & Robusto, 2007; Kubinger, 2009). Thus, item explanatory IRT models using the LLTM approach can provide useful and practical information for improving test development, item generation, and educational assessments.

Despite such potential uses in educational measurement practices, it appears that item explanatory models have been less commonly used for polytomous data than for dichotomous data. Most of the LLTM applications to item explanatory measurement have been in the context of dichotomous data (e.g. Fischer, 1973; Hartig, Frey, Nold, & Klieme, 2012; Kubinger, 2009; Poinstingl, 2009; Sonnleitner, 2008). However, this approach needs to be further investigated for polytomous data, because it is very common to have polytomous data such as ordered-category

responses in a wide range of educational, psychological, and sociological applications (De Boeck & Wilson, 2004). Although the LLTM itself was originally developed for dichotomous data by extending the simple Rasch model (Fischer, 1973), its general approach to item explanatory measurement, which decomposes the difficulties of specific items into linear combinations of elementary components, can also be applied to polytomous data.

The few studies that have investigated polytomous extensions of the LLTM approach include Fischer and Parzer (1991), Fischer and Ponocny (1994), Glas and Verhelst (1989), and Linacre (1989). Fischer and Parzer (1991) applied the LLTM approach to the rating scale model and later Fischer and Ponocny (1994) applied it to the partial credit model. However, these applications are based on conditional maximum likelihood estimation, and their item parameterization with a normalization constant and basic parameter is complicated and less interpretable. Glas and Verhelst (1989) imposed linear restrictions on the item parameters of the partial credit model using marginal maximum likelihood estimation, but their approach necessarily requires a translation of the estimated item parameters and hence it is hard to interpret them directly. Meanwhile, Linacre (1989) proposed the Many-Facet Rasch Model (MFRM) using several maximum likelihood estimation methods. The MFRM is an extension of the LLTM to polytomous items (Wu, Adams, Wilson & Haldane, 2007), and it is useful and flexible to form a measurement model which reflects various measurement situations such as examinee, items, raters, tasks, interviewers, or scoring criteria (Eckes, 2009; Linacre, 1989). Despite the usefulness and flexibility of the MFRM, it has been mainly used to investigate the effects of facets[1] such as rater effects (e.g. Bonk & Ockey, 2003; Draney & Wilson, 2008; Eckes, 2009). However, the MFRM approach can be used for item explanatory measurement to examine the effects of item properties on the difficulties of polytomous items.

The aim of this study is to demonstrate how item explanatory models using the MFRM approach, regarded as a polytomous extension of the LLTM approach, are applied to polytomous data as well as to investigate how the observed item properties explain and predict the overall item difficulties of polytomous items empirically. For empirical examples, I will use the Carbon Cycle assessment data (Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009) and the Social Evaluative Reasoning (SER) in the workplace data (Jolin, 2015, 2018) as well.

The Carbon Cycle assessment was developed based on a learning progression for carbon cycling in socio-ecological systems in science education. The test construction of the learning progression has four ordered levels of achievement corresponding to students' progress toward more sophisticated reasoning about biogeochemical processes (Mohan et al., 2009). The Carbon Cycle items were scored as ordered-category responses, and they consisted of a combination of three categorical item properties: (a) three types of biogeochemical processes in carbon cycling (cellular respiration, photosynthesis, digestion/biosynthesis), (b) four learning progress variables in which carbon cycling processes (large-scale systems, micro-scale systems, energy, mass), and (c) two item response formats (multiple choice with explanation, yes/no choice with explanation).

The SER in the workplace data were derived from an instrument designed as a measure of soft skill proficiency for transition-age secondary students with disabilities. In total, 16 workplace scenarios in a visual narrative format depict a target employee responding either correctly or

[1] Widely defined, a facet can be any factor, variable, or component of the measurement situation that is assumed to influence test scores in a systematic way (Eckes, 2009; Linacre, 1989). Individual items and item properties, i.e., unique characteristics inherent in a set of items, can be considered as an item facet.

incorrectly—according to the dictates of service-sector employment—to the social cues emoted by customers. Within each scenario, Evaluative Inference (EI) ability items were scored as three ordered levels of evaluative inferencing ability based on the inferencing taxonomy by Pearson (1978). Social Perceptual Unit (SPU) detection ability items were scored as five ordered levels. The complexity of the item construct was manipulated on the basis of the frequency and co-occurrence of three factors of item properties: (a) types of emotional cues present in the scenarios: basic versus complex (Baron-Cohen et al., 2001), (b) types of language used to deliver social cues: figurative versus literal (Happe, 1995), and (c) outcome resolution by the target employee (i.e., correct versus incorrect resolution).

The items in these empirical data sets were newly developed with a set of item properties in their item design or test construction. Item explanatory models using the MFRM approach are necessary and helpful to figure out the effects of those item properties on the overall item difficulties. Moreover, by composing specified combinations of the item property effects, we can predict the overall item difficulty for developing items in a scientific and systematic way. Through the two empirical studies, applications of the MFRM approach for item explanatory measurement will be demonstrated to show how the MFRM-based item explanatory models work in practice and how item properties explain and predict the overall item difficulties of polytomous items.

2.2. Item Explanatory IRT Models under the Multidimensional Random Coefficient Multinomial Logit Model Framework

2.2.1. The Multidimensional Random Coefficient Multinomial Logit Model Framework

The Multidimensional Random Coefficients Multinomial Logit (MRCML; Adams, Wilson, & Wang, 1997) model framework is used to implement all item response models in this paper. The MRCML model is a very general framework, the models of which share fundamental measurement properties of the Rasch family models (Adams, Wilson, & Wu, 1997). It covers a wide range of the Rasch family models including the simple Rasch model, the Rating Scale Model (Andrich, 1978), the Partial Credit Model (Masters, 1982), the Linear Logistic Test Model (Fischer, 1973), the Many-Facet Rasch Model (Linacre, 1989), their multidimensional forms, their latent regression extensions, and others designed for more complex measurement features and situations (Briggs & Wilson, 2003). Moreover, it affords considerable flexibility to allow the design of customized models for specific test situations (Adams, Wilson, & Wang, 1997; Wilson, Zheng, & McGuire, 2012). In terms of generalizability and flexibility in item response modeling, it is attractive to use the MRCML model framework.

In particular, considering item explanatory models which impose linear restrictions on the item parameters, the extended Rasch family models for item explanatory models under the MRCML model framework can be illustrated as in Figure 2-1. The LLTM is a typical item explanatory model for dichotomous items, and the MFRM is a flexible item explanatory model for polytomous items including partial credit items and rating scale items.

Under the MRCML model framework, the probability that person p with the D-dimensional ability vector $\boldsymbol{\theta}_p$ will select or generate a response in category m ($m = 0, 1, \ldots, M_i$) to item i ($i = 1,2, \ldots, I$) is modeled as

$$Pr\left(Y_{pi} = m; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} \middle| \boldsymbol{\theta}_p\right) = \frac{\exp\left(\mathbf{b}_i' \boldsymbol{\theta}_p + \mathbf{a}_{im}' \boldsymbol{\xi}\right)}{\sum_{m=1}^{M_i} \exp\left(\mathbf{b}_i' \boldsymbol{\theta}_p + \mathbf{a}_{im}' \boldsymbol{\xi}\right)} \ , \tag{2.1}$$

where $\boldsymbol{\theta}_p$ is a $D \times 1$ column vector of D-dimensional latent abilities for person p ($\boldsymbol{\theta}_p = \theta_{1p}, \theta_{2p}, \ldots, \theta_{Dp}$), $\mathbf{A}$ is a design matrix of the design vectors $\mathbf{a}_{im}$ ($\mathbf{a}_{im}$ is a row of $\mathbf{A}$), $\mathbf{B}$ is a scoring matrix of the scoring vectors $\mathbf{b}_i$ ($\mathbf{b}_i'$ is a row of $\mathbf{B}$), and $\boldsymbol{\xi}$ is a vector of item parameters. Note that for model identification a response in category 0 is denoted by a vector of 0s, which makes the 0 category a reference category (Adams, Wilson, & Wang, 1997). In the MRCML model framework, the design and scoring matrices are used to specify the functional form of the Rasch family models (e.g. Adams & Wilson, 1996; Briggs & Wilson, 2003; Wilson et al, 2012). For example, for the Rating Scale Model in which all items have the same number of categories ($M_i = M$), $\mathbf{A}$ is matrix of $I \times M$ rows with $\mathbf{A} = [\mathbf{a}_{11}, \mathbf{a}_{12}, \ldots, \mathbf{a}_{1M}, \ldots, \mathbf{a}_{I1}, \mathbf{a}_{I2}, \ldots \mathbf{a}_{IM}]'$, $\mathbf{B} = [\mathbf{b}_1', \mathbf{b}_2', \ldots, \mathbf{b}_I']'$ where $\mathbf{b}_i = [b_{i1}, b_{i2}, \ldots, b_{iD}]'$, and $\boldsymbol{\xi} = (\delta_1, \delta_2, \ldots, \delta_i, \tau_1, \tau_2, \ldots, \tau_{M-1})$. In MRCML models, the item parameters and person parameters are estimated by using maximum likelihood estimation methods such as conditional maximum likelihood and marginal maximum likelihood (Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003).

ConQuest software (Adams, Wu & Wilson, 2012; Wu et al., 2007) has been designed to implement the MRCML model framework using marginal maximum likelihood estimation. It can fit MRCML models using any estimable and identifiable pair of item design and scoring matrices as well as generate those matrices from a user-friendly model-building language (Wilson, Zheng, & McGuire, 2012).

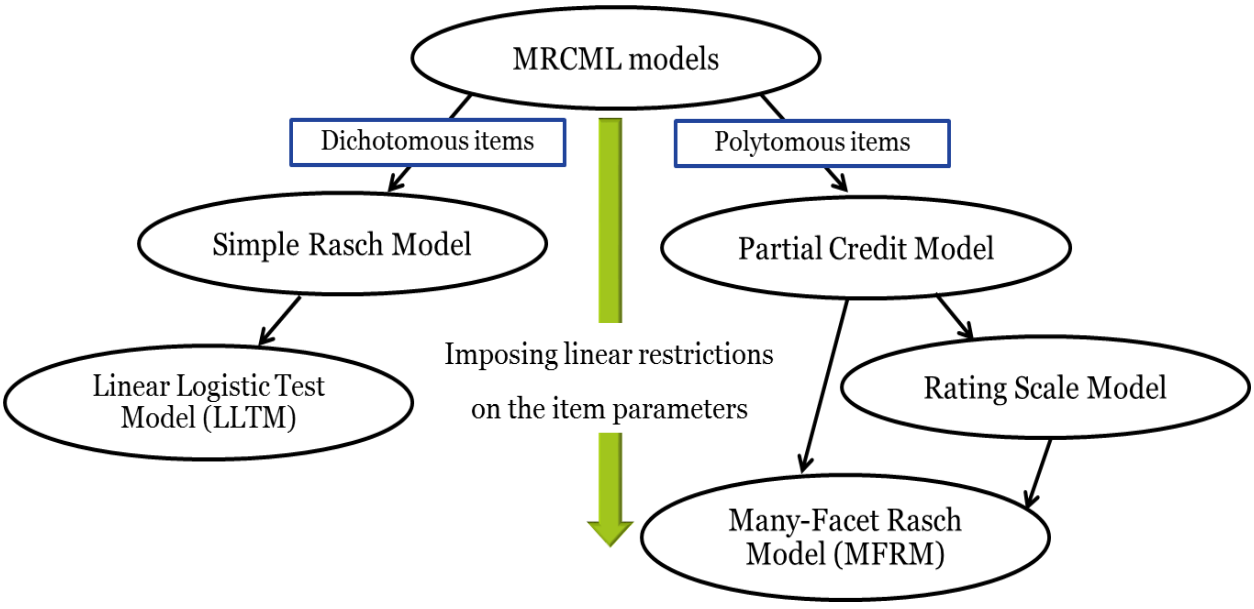


Figure 2-1. Item explanatory models under the MRCML model framework

2.2.2. Item Response Models for Polytomous data

The focus of this study is to demonstrate applications of item explanatory models using the MFRM approach to polytomous data. The two empirical data sets, the Carbon Cycle assessment data and the SER in the workplace data, used in this paper have ordered-category responses as described briefly in the introduction section. Given the ordered-category responses regarded as polytomous data, the adjacent-categories logit relationship is employed for polytomous Rasch family models including item explanatory models. Adjacent-categories logit based Rasch family models can make person and item parameters separable as well as enable specifically objective comparisons of persons and items (Masters, 1982). Adjacent-categories logits are suitable for subjectively assigned scores between categories in ordered responses (Anderson, 1984; Johnson, 2007), and they are useful for local comparison when contrasting probabilities of the responses in pairs of ordered adjacent categories (Hartzel, Agresti, & Caffo, 2001; Masters & Wright, 1997).

Under the MRCML model framework, the Partial Credit Model (PCM; Masters, 1982) and the Rating Scale Model (RSM; Andrich, 1978) are representative polytomous Rasch family models based on the ordered-category responses and adjacent-categories logits. In the PCM, the conditional probability that person p with ability θ_p will respond with category score m on item i with step difficulty δ_{im} is modeled as:

$$Pr\left(y_{pi} = m \middle| \theta_p\right) = \frac{\exp\left(\sum_{l=0}^{m}(\theta_p - \delta_{il})\right)}{\sum_{h=0}^{M_i}\exp\left(\sum_{l=0}^{h}(\theta_p - \delta_{il})\right)}\,, \qquad m = 0, 1, \dots, M_i\,, \tag{2.2}$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$, $\delta_{i0} = 0$, and $\sum_{l=0}^{0}(\theta_p - \delta_{il}) = 0$ so that $\sum_{l=0}^{h}(\theta_p - \delta_{il}) = \sum_{l=1}^{h}(\theta_p - \delta_{il})$. Note that the mean person ability is constrained to zero for model identification[2]. This adjacent category logit model is a special case of a multinomial logit model (De Boeck & Wilson, 2004; Glas & Verhelst, 1989). In a more effective way of interpreting polytomous item parameters, the PCM is described in terms of local comparison in the response probabilities from the adjacent category score $m-1$ to m as follows:

$$ln\frac{Pr\left(y_{pi} = m \middle| \theta_p\right)}{Pr\left(y_{pi} = m-1 \middle| \theta_p\right)} = \theta_p - \delta_{im}\,, \tag{2.3}$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$ and $\delta_{i0} = 0$. From this local comparison perspective of the PCM, the item parameter δ_{im} is interpreted as a "step" difficulty for scoring m rather than $m-1$ on item i ($i = 1,2, \dots, \mathrm{I}$), which is regarded as the m-th step of switching over to the next response category. Thus, the step difficulty parameter δ_{im} indicates the relative difficulty when shifting the category score from $m-1$ to m on item i (Embretson & Reise, 2000; Masters, 1982; Zheng & Rabe-Hesketh, 2007).

[2] For model identification, a particular item parameter or the mean of item parameters is constrained to zero, instead of constraining the mean person ability (De Boeck & Wilson, 2004).

Interestingly, the ConQuest software (Adams, Wu & Wilson, 2012; Wu et al., 2007) uses a twofold item parameterization, which divides the step difficulty parameter δ_{im} into two item parameters of the item location parameter β_i and the step deviation parameter τ_{im} as follows:

$$\delta_{im} = \beta_i + \tau_{im}, \tag{2.4}$$

so that

$$ln\frac{Pr(y_{pi} = m|\theta_p)}{Pr(y_{pi} = m-1|\theta_p)} = \theta_p - \beta_i - \tau_{im}\,, \qquad m = 1, \ldots, M_i\,, \tag{2.5}$$

where $\theta_p \sim N(0, \sigma_\theta^2)$, $\tau_{i0} = 0$, and $\sum_{m=1}^{M_i} \tau_{im} = 0$ so that $\frac{1}{M_i}\sum_{m=1}^{M_i} \delta_{im} = \beta_i$. Here, β_i is an item location (a.k.a. overall item difficulty) parameter for item i, and τ_{im} is a step deviation parameter for the m-th step within item i. Note that the item location parameter is interpreted as the overall item difficulty for each polytomous item due to the model constraint that $\frac{1}{M_i}\sum_{m=1}^{M_i} \delta_{im} = \beta_i$. The step deviation parameter τ_{im} differs from the step difficulty parameter δ_{im} in that the step deviation is a deviation from the item location to the step difficulty ($\tau_{im} = \delta_{im} - \beta_i$). For model identification, the sum of step deviations is constrained to be zero for each item ($\sum_{m=1}^{M_i} \tau_{im} = 0$), so that $M_i - 1$ step deviation parameters are estimated for each item. The twofold item parameterization is useful to recognize item locations, scale thresholds, and step difficulties in polytomous items as well as to incorporate item properties to explain such item characteristics. It is also helpful to constrain item parameters for model identification.

The RSM considers rating scale items which are ordinally scored to represent varying degrees of the construct. It is assumed that the relative difficulties of the steps between categories are constant for all items ($\tau_{im} = \tau_m$) and the number of steps are the same for all items ($M_i = M$). In the RSM, the item location parameters β_i and the common scale threshold parameters τ_m can be expressed as:

$$ln\frac{Pr(y_{pi} = m|\theta_p)}{Pr(y_{pi} = m-1|\theta_p)} = \theta_p - \beta_i - \tau_m\,, \qquad m = 1, \ldots, M\,, \tag{2.6}$$

where $\theta_p \sim N(0, \sigma_\theta^2)$, $\tau_0 = 0$, and $\sum_{m=1}^{M} \tau_m = 0$. The RSM is a special case of the PCM because it imposes more restrictions on the item parameters. In other words, the PCM is a generalization of the RSM (Mair & Hatzinger, 2007).

2.2.3. Item Explanatory Models Using the Many-Facet Rasch Model

In order to apply item explanatory models to polytomous data under the MRCML model framework, the MFRM can be used to examine the effects of item properties on the overall item difficulties of polytomous items. Although the MFRM is regarded as a polytomous extension of the LLTM, model notations and types of item predictor variables are the main differences between the two approaches. While the LLTM uses general regression notation so that it can incorporate both categorical and continuous predictors, the MFRM is typically expressed using factorial

(ANOVA-style) notations so that it typically uses categorical predictors. Conceptually and practically, the MFRM is an extensive and flexible measurement model because it can take into account various measurement situations or facets such as examinee, items, raters and tasks (Eckes, 2009; Linacre, 1989).

To specify item explanatory models using the MFRM approach, categorical item properties are considered as facets or sub-facets within the item facet (Wang & Wilson, 2005). For a simple model specification, suppose that three categorical item properties are incorporated in the RSM. Considered as an item explanatory model, each of the original items is modeled as a new generalized item which is constructed as a linear combination of the three item properties. The subscript i for the original item in Equation 2.6 will now be substituted for three-factorial subscript hjk for the new generalized item. Also, the item location parameter β_i will be decomposed into the effects of the three item properties $h,\ j$ and k, which are categorical predictors in the design matrix under the MRCML model framework. These factor notations extend to any number of item properties but become more complex as the number of them increases.

In terms of local comparison in the response probabilities from the adjacent category score $m-1$ to m, the MFRM is expressed as:

$$ln\frac{Pr\left(y_{phjk}=m\middle|\theta_p\right)}{Pr\left(y_{phjk}=m-1\middle|\theta_p\right)}=\theta_p-\alpha_h-\gamma_j-\omega_k-\tau_m\ ,\qquad m=1,\ldots,M\ , \tag{2.7}$$

where $\theta_p \sim N\left(0,\sigma_\theta^2\right)$, $\sum_{j=1}^{J}\gamma_j=\sum_{k=1}^{K}\omega_k=0$[3], $\tau_0=0$, and $\sum_{m=1}^{M}\tau_m=0$. Here, θ_p is the latent ability of person p, α_h is the effect of item property $h\ (h=1,2,\ldots,H)$ on the overall item difficulties, γ_j is the effect of item property $j\ (j=1,2,\ldots,J)$ on the overall item difficulties, ω_k is the effect of item property $k\ (k=1,2,\ldots,K)$ on the overall item difficulties, and τ_m is a scale threshold parameter which represents the added difficulty of stepping up from the adjacent category score $m-1$ to m. By applying the MFRM approach to the RSM, item property factors can account for the overall difficulties of polytomous items. This is equivalent to Fischer and Parzer (1991)'s model, referred to as the linear rating scale model, when categorical item properties are used. In this form of the MFRM, the item property effects are additive without interactions between them.

In addition, since the MFRM is flexible to include an interaction term between factors as in ANOVA, one can take into account the interaction effects of item properties as well as the main effects of them (see Eckes, 2009; Linacre, 1989). If an interaction between the item property *j* and the item property k is considered, the modified model with adding an interaction term is formed as:

$$ln\frac{Pr\left(y_{phjk}=m\middle|\theta_p\right)}{Pr\left(y_{phjk}=m-1\middle|\theta_p\right)}=\theta_p-\alpha_h-\gamma_j-\omega_k-\varphi_{jk}-\tau_m\ ,\qquad m=1,\ldots,M\ , \tag{2.8}$$

[3] The MFRM typically uses effect coding for categorical item properties so that the effect of each item property is constrained to zero sum (see Wu et al., 2007). Here, due to the model constraint of zero mean person ability, the effect of the first item property α_h is not constrained. Otherwise, it is constrained that $\sum_{h=1}^{H}\alpha_h=0$.

where $\sum_{j=1}^{J}\sum_{k=1}^{K}\varphi_{jk}=0$, φ_{jk} is the interaction effect between the item property j and the item property k, and all other parameters are the same in Equation 2.7. In terms of goodness-of-fit, this MFRM form with an interaction may fit the data better due to richer predictors.

On the other hand, when applying the MFRM approach to the PCM, the step deviation parameter τ_{im} in Equation 2.5 can be estimated flexibly and variously according to the scoring scale structure to meet the requirements of specific measurement situations or assessment contexts (see Eckes, 2009; Linacre, 1989). For instance, if the scale to score partial credit items is determined by all three item properties, the step deviation parameter will be estimated for the new generalized item, with three-factorial subscript hjk, constructed as a linear combination of those item properties as follows:

$$ln\frac{Pr\left(y_{phjk}=m\middle|\theta_p\right)}{Pr\left(y_{phjk}=m-1\middle|\theta_p\right)}=\theta_p-\alpha_h-\gamma_j-\omega_k-\tau_{hjkm}\,,\qquad m=1,\ldots,M_{hjk}\,, \tag{2.9}$$

where $\theta_p\sim N\left(0,\sigma_\theta^2\right)$, $\sum_{j=1}^{J}\gamma_j=\sum_{k=1}^{K}\omega_k=0$, $\tau_{hjk0}=0$, and $\sum_{m=1}^{M_{hjk}}\tau_{hjkm}=0$, and also where$\tau_{hjkm}$ is a step deviation parameter for the new generalized item constructed by the three item properties h, j and k when shifting the category score from $m-1$ to m, and other parameters for the item property effects (α_h, γ_j, and ω_k) are defined as before in Equation 2.7. In this MFRM variation, although the scoring scale structure and the estimated step deviation parameter can vary, the model has the same item explanatory component in which the overall difficulties of polytomous items are decomposed into linear combinations of the item property effects. Thus, the MFRM approach under the MRCML model framework has the flexibility and usefulness in modeling item explanatory measurement, which goes beyond descriptive measurement concerning individual differences of elements of the measurement.

2.3. Empirical Studies

2.3.1. Empirical Data Analysis

Two empirical studies were conducted to demonstrate how item explanatory models using the MFRM approach can be applied to polytomous data under the MRCML model framework. For empirical data sets, the Carbon Cycle assessment data and the Social Evaluative Reasoning (SER) in the workplace data were used in each study to examine how their observed item properties explain and predict the overall item difficulties in practice.

In order to find an item explanatory model that provides the most parsimonious fit to the data, I compared several MFRM-based item explanatory models which incorporate additive effects of the item properties with or without their interaction effects. In addition, to compare the performance of those item explanatory models with a general polytomous Rasch model, the PCM was fitted to the data. The PCM was used as a reference model because it is a saturated model with no linear restrictions on the item parameters in polytomous Rasch family models under the MRCML model framework, as shown in Figure 2-1.

For data analyses, the ConQuest 3.0 software (Adams, Wu & Wilson, 2012), which uses marginal maximum likelihood (MML) estimation, was used to fit all the adjacent-categories logit based item response models under the MRCML model framework. For the purpose of model identification, the mean person ability was constrained to zero for all the fitted models. By default, in the ConQuest software, effect coding was used to incorporate categorical predictors for the categorical item properties into the design matrix under the MRCML model framework (see Wu et al., 2007). In addition, R software (R Development Core Team, 2008) was also used for supporting statistical analyses and creating plots for graphical model comparisons.

To evaluate the performance of the MFRM-based item explanatory models, two approaches were employed for model comparisons: comparison of goodness-of-fit and comparison of parameter estimates. For the first approach, three goodness-of-fit indices were reported for each model. The deviance, the Akaike Information Criterion (AIC; Akaike, 1974), and the Bayesian Information Criterion (BIC; Schwarz, 1978) are computed as:

$$Deviance = -2 \times Ln(Likelihood),$$
$$\text{AIC} = Deviance + 2k,$$
$$\text{BIC} = Deviance + Ln(N) \times k,$$

where *k* is the number of estimated parameters and *N* is the number of observations. The AIC and BIC are computed from the deviance with a penalty included for the number of estimated parameters so that they can be viewed as a measure that combines model fit (accuracy) and complexity (parsimony) of the fitted model (Kuha, 2004). When candidate models are fitted to the same data, smaller values of these indices indicate better goodness-of-fit of the model.

The Likelihood Ratio test (LR test; Andersen, 1973) was also used to compare goodness-of-fit between nested models. The LR test is often used to examine the model fit of item explanatory models compared with the saturated model (see De Boeck & Wilson, 2004; Poinstingl, 2009). For example, the likelihood of the MFRM (L_{MFRM}) is compared to that of the PCM (L_{PCM}), so that $-2 \times Log\left(\frac{L_{MFRM}}{L_{PCM}}\right) = Deviance_{MFRM} - Deviance_{PCM}$ is asymptotically distributed as a χ^2 with a number of degrees of freedom (*df*) equal to the difference between the number of parameters of the two models. The 5% level of significance was used for all of the statistical significance tests.

In addition, to see if there is agreement for the estimated and calculated step difficulties between the fitted models, a graphical comparison and a correlation analysis were conducted (see Fischer, 1973; Poinstingl, 2009). The step difficulties were calculated using the estimated item parameters in the MFRM-based item explanatory models, and then those values were compared with the corresponding step difficulties which were estimated in the PCM.

In each empirical study, the item property effects found from the best fitting item explanatory model were examined to see how the observed item properties explain and predict the overall item difficulties in practice. Although both the item property effect parameters and the step deviation (or scale threshold) parameters are estimated in the MFRM approach, the former parameters are matters of concern for interpreting the results. In the MFRM-based item explanatory models, the item location parameters representing the overall item difficulties are the target difficulty parameters of polytomous items which are to be meaningfully explained by the item properties. Hence, the effects of item properties on the overall item difficulties are the primary interest of this research.

2.3.2. Application to the Carbon Cycle Assessment data

Data

In the first empirical study, a subset of the Carbon Cycle assessment data was used, which was collected from the pre-test of the Math and Science Partnership (MSP) Carbon student assessment in the 2010-2011 academic year. Middle and high school students from urban, suburban, and rural area in five states (CA, CO, MD, MI, NY) participated in the multi-year assessments. In the Carbon Cycle project for science education, a learning progression for carbon cycling in socio-ecological systems was conceptually coherent and empirically validated by an iterative process of designing and analyzing the assessments and interviews for secondary school students (Mohan et al., 2009). The assessments and items were developed based on the learning progression in which the research-based construct has four ordered levels of achievement corresponding to students' progress toward more sophisticated reasoning about biogeochemical processes (Jin & Anderson, 2012; Mohan et al., 2009), as in Table 2-1.

Students at Level 1 tend to make explanations limited to macroscopic processes about organisms and objects in terms of the action-result chain, and they tend to use everyday language rather than technical vocabulary or scientific representations. Level 2 students tend to explain macroscopic processes in terms of materials changed by hidden mechanisms or unobservable actors, but they focus on actors and results rather than changes. Students at Level 3 tend to recognize the transformation of matter and explain macroscopic processes in terms of changes of molecules and energy forms or chemical changes, but their explanations are restricted due to their lack of understanding of chemical substances. Level 4 students tend to identify the cell as the basic unit and use atomic-molecular ideas and chemical models consistently to explain macroscopic processes by linking carbon-transforming processes with matter and energy as constraints. They trace matter through hierarchically organized systems to explain changes at different scales systematically.

Table 2-1. The Learning Progression Framework for the Carbon Cycle items

Level	General description	Exemplary response
4	Linking processes with matter and energy as constraints	The mass decrease because the mold uses energy to grow it goes through cellular respiration, the result of cellular respiration is CO2 and H20
3	Changes of molecules and energy forms with unsuccessful constraints	The mold decays the bread and turns it into energy so the mass decreases
2	Force-dynamic accounts with hidden mechanisms	Even though the mold is converting the bread into biomass, the mold also has gained mass through photosynthesis.
1	Macroscopic force-dynamic accounts	The mass increases because the mold adds its weight to the original mass of the bread

The construct map for the Carbon Cycle items has core characteristics and indicators for each level, derived from students' responses to open-ended questions and interviews. The items asked students to answer forced choice questions and explain their choices, and their responses were scored ordinally according to scoring rubrics of the achievement levels based on the learning progression. For example, the scoring rubric of the item 1 is represented in Figure 2-2. Four levels are assigned to score student achievement according to the scoring rubric. Students in the highest level are hypothesized to be able to correctly link the carbon-transforming processes of cellular respiration with energy and matter as constraints. Students' responses reaching this scoring criterion are scored at Level 4. Similarly, this scoring rubric works for the other levels. If students' answers are not related to the question, or are illegible or nonsensical, they are coded as missing values.

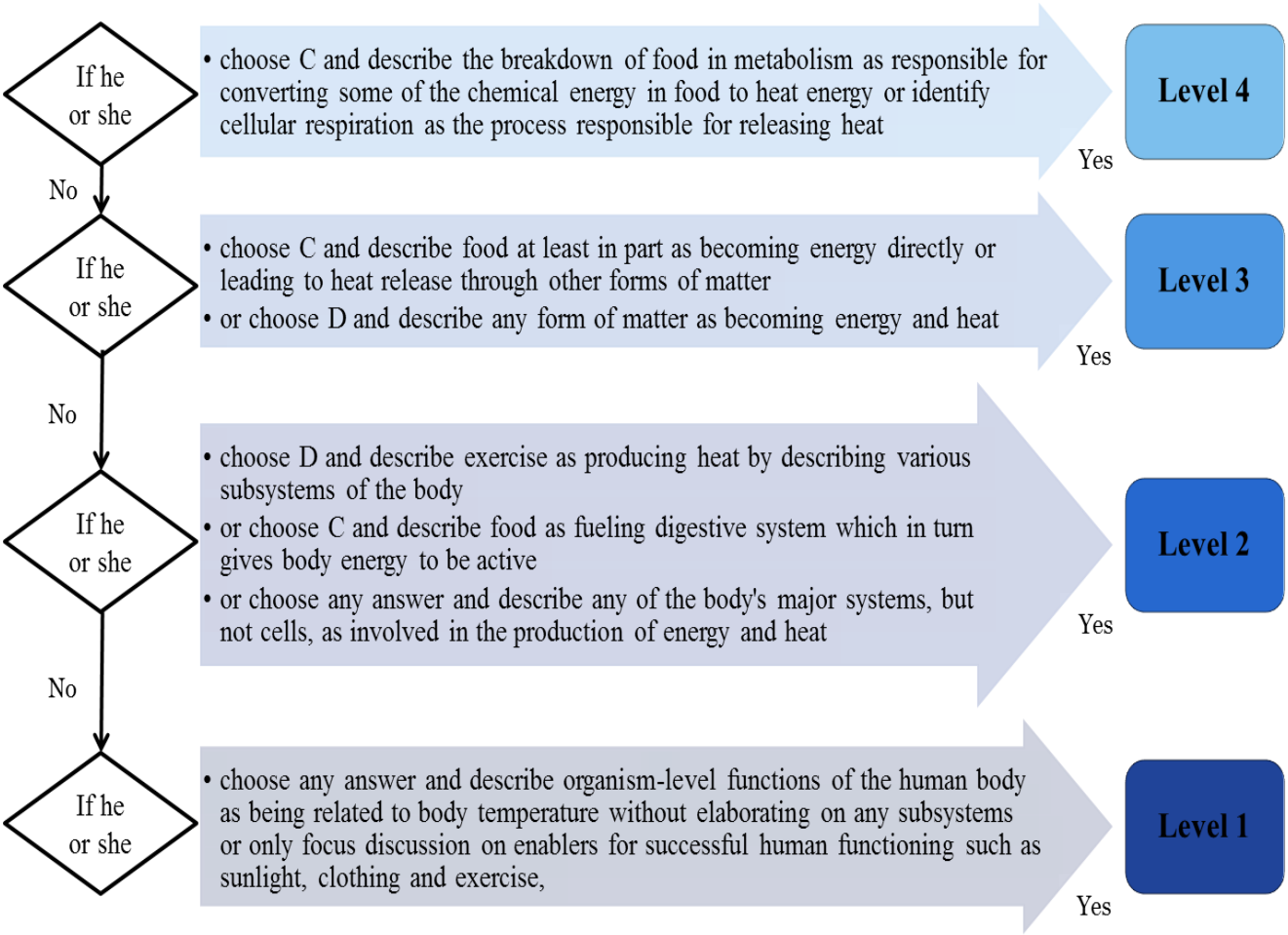


Figure 2-2. Flow chart of the scoring rubric for the item 1 (BODYTEMP)

As in Table 2-2, the 13 Carbon Cycle items consist of three categorical item properties, regarded as facets or sub-facets within the item facet. The three item properties determining the item locations are embedded in the design of the Carbon Cycle items. The first one is the Process property that concerns biogeochemical processes in carbon cycling that transform carbon in socio-ecological systems at multiple scales (Mohan, Chen, & Anderson, 2009). These processes are oxidizing organic carbon (cellular respiration; CR), generating organic carbon (photosynthesis;

PS), and transforming organic carbon (digestion/biosynthesis; DB). Second, the Progress property represents progress variables in which the carbon cycle processes occur: large-scale systems (LS), micro-scale systems (MS), energy (EN), and mass (MA). The third property is the Format of items distinguishing between multiple choice with explanation (MC) and yes/no choice with explanation (YN). Predictors for the item properties are working as weights of the elementary components gathered into a Q matrix in the LLTM approach (Kubinger, 2009; Poinstingl, 2009).

Each item can be identified as a new generalized item constructed by the three item properties. For instance, the item 1 (BODYTEMP) can be identified as a combination of a CR predictor (Process property), a EN predictor (Progress property) and a MC predictor (Format property). The text of item 1 is shown below:

Q1. [BODYTEMP] Your body produces heat to maintain its normal temperature. Where does the heat mainly come from? Please choose the ONE answer that you think is best and explain why you think that the answer you chose is better than the others. If you think some of the other answers are also partially right, please explain why you think so.

A. The heat mainly comes from sunlight.
B. The heat mainly comes from the clothes you are wearing.
C. The heat mainly comes from the foods you eat.
D. The heat mainly comes from your body when you are exercising.

Table 2-2. Q matrix for the Carbon Cycle items composed of three item properties

Item		Process			Progress				Format	
Num.	Name	Cellular Respiration (CR)	Photosyn-thesis (PS)	Digestion / Biosynthesis (DB)	Large Scale (LS)	Micro Scale (MS)	Energy (EN)	Mass (MA)	Multiple Choice (MC)	Yes/No Choice (YN)
1	BODYTEMP	1	0	0	0	0	1	0	1	0
2	CARBCYC2	1	0	0	1	0	0	0	0	1
3	CARBPLNT	0	1	0	0	1	0	0	0	1
4	CONTCARB	0	0	1	0	1	0	0	0	1
5	CRKTGRWTH	1	0	0	0	0	0	1	1	0
6	EATAPPLE	0	0	1	0	1	0	0	0	1
7	INFANT	0	0	1	0	0	0	1	0	1
8	LIGHTEN	0	1	0	0	0	1	0	1	0
9	LIGHTEN2	0	1	0	0	0	1	0	1	0
10	MAPLEMASS	0	1	0	0	0	0	1	1	0
11	PLNTGSENS	1	0	0	0	1	0	0	0	1
12	PLNTGRWTH	0	1	0	0	0	0	1	0	1
13	THINGTREE	0	1	0	0	0	0	1	0	1

In the initial data, there were some items which had too few or zero responses for the highest level, so that they resulted in poor or unfeasible estimation of the third step's difficulty parameters (e.g. CARBCYC2, CONTCARB, CRKTGRWTH, MAPLEMASS, PLNTGSENS, THINGTREE).

To avoid such sparse response problem, item responses of the four level categories (1, 2, 3, 4) were recoded into three category scores (0, 1, 2, 2) in this empirical study. Thus, we can interpret the second step's difficulties as the added difficulties as one goes from the Level 2 to the combined Levels of 3 and 4. In addition, some cases which have less than three valid item responses were deleted to enhance the quality of estimation and reduce the error of estimation for person abilities. After refining the data, in total, the three levels of ordered-category responses from 1,157 students answered to the 13 Carbon Cycle items and the three categorical item properties were included in the final sample for data analyses.

Results

According to the data analysis procedure, the PCM and item explanatory models using the MFRM approach were fitted to the Carbon Cycle assessment data using the MRCML model framework. Two item explanatory models were compared to find the most parsimonious fitting model: a MFRM with additive item property effects (called the additive MFRM here) and a MFRM with additive item property effects and an interaction between the Process and Progress properties (called the interaction MFRM here). Based on the learning progression framework for the Carbon Cycle items, the scale structure was determined by the three item properties in the two MFRM-based item explanatory models.

Table 2-3. Goodness-of-fit of item explanatory models to the Carbon Cycle assessment data

Models		PCM	Item explanatory models	
			additive MFRM	interaction* MFRM
Good-ness of Fit	Deviance	12223.58	13021.69	12828.39
	AIC	12277.58	13057.69	12876.39
	BIC	12414.03	13148.65	12997.68
	k	27	18	24

Note: k = number of estimated parameters; * an interaction between the Process and Progress properties.

Table 2-3 shows the goodness-of-fit results of the fitted models to see performance of the two item explanatory models. In terms of goodness-of-fit indices, the AIC and BIC, both item explanatory models appeared to fit worse than the PCM, as expected, that is item explanatory models commonly fit worse than the saturated model (Kubinger, 2009). This result was confirmed in that the LR test comparing with the PCM was significant for the additive MFRM ($\chi^2(9) = 798.11$, $p < .001$) as well as for the interaction MFRM ($\chi^2(3) = 604.81$, $p < .001$). Moreover, 27 parameters estimated in the PCM were reduced to 18 in the additive MFRM and 24 in the interaction MFRM. A methodological advantage of a smaller number of estimated item parameters in item explanatory models is at the cost of lower goodness-of-fit (De Boeck & Wilson, 2004). When comparing between the two item explanatory models, the interaction MFRM fitted better than the additive MFRM. The LR test between them was significant as $\chi^2(6) = 193.30$ ($p < .001$), and both the AIC and BIC indices showed a superior goodness-of-fit for the interaction MFRM.

Table 2-4. Step difficulties for the Carbon Cycle items in item explanatory models

Item	Step difficulty	PCM	additive MFRM	interaction MFRM
1	δ_{11}	0.23	-0.02	0.01
	δ_{12}	2.16	1.54	1.83
2	δ_{21}	-4.02	-4.06	-4.16
	δ_{22}	2.90	2.60	2.59
3	δ_{31}	-1.78	-1.84	-1.73
	δ_{32}	1.55	1.31	1.47
4	δ_{41}	-1.64	-1.59	-1.96
	δ_{42}	1.79	1.48	1.28
5	δ_{51}	-0.37	-0.30	-0.58
	δ_{52}	2.39	3.21	2.07
6	δ_{61}	-2.61	-1.59	-1.96
	δ_{62}	0.93	1.48	1.28
7	δ_{71}	-0.59	-0.70	-0.32
	δ_{72}	2.81	1.88	2.91
8	δ_{81}	-0.76	-0.56	-0.67
	δ_{82}	0.77	0.42	0.41
9	δ_{91}	-0.66	-0.56	-0.67
	δ_{92}	0.20	0.42	0.41
10	δ_{101}	0.94	0.55	0.71
	δ_{102}	1.57	0.71	1.24
11	δ_{111}	-0.44	-0.41	-0.18
	δ_{112}	1.87	1.52	1.97
12	δ_{121}	0.81	-0.14	-0.31
	δ_{122}	2.33	0.90	0.76
13	δ_{131}	-2.51	-0.14	-0.31
	δ_{132}	0.03	0.90	0.76

In addition to the statistical goodness-of-fit comparison, effect size was also investigated: agreement for the estimated and calculated step difficulties between the fitted models was examined by a graphical comparison, and a correlation analysis was used to see which gave the better calculated step difficulties. In Table 2-4, the step difficulties for the Carbon Cycle items were calculated by the estimated item parameters in the two MFRM-based item explanatory models and they were freely estimated in the PCM. Compared to the step difficulties for the original items in the PCM, corresponding values for the new generalized items in each item explanatory model were compared graphically as shown in Figure 2-3.

A graphical comparison shows that the interaction MFRM had similar agreement to the additive MFRM although a few more step difficulty points in the interaction MFRM were located close to the 45-degree line indicating perfect alignment. The second step difficulties of item 5, 7 and 10 were closer to the 45-degree line in the interaction MFRM than in the additive MFRM. Yet,

most of the step difficulties were similarly located and the first step difficulty of item 13 were located farthest from the 45-degree line in the two models. These results were confirmed by a correlation analysis. The correlation of step difficulties with the PCM was 0.91 for the additive MFRM, whereas it was 0.93 for the interaction MFRM. In addition, the proportion of the total step difficulty variance in the PCM that is explained by the MFRM-based item explanatory model (R^2) was calculated. The R^2 value was 0.72 for the additive MFRM and 0.78 for the interaction MFRM, which means that the interaction MFRM accounted for variability in the step difficulties of the PCM more by 6% than the additive MFRM.

Consequently, the interaction MFRM performed better than the additive MFRM in terms of the goodness-of-fit and the explained variability in the step difficulties, although the two models revealed comparable performance in reconstructing the step difficulties using the item property effects. This makes sense because the interaction MFRM had more predictors to explain the item effects as well as the scale structure. Since the interaction MFRM showed the superior performance in model comparison, it was determined as the final item explanatory model to investigate how the three item properties (Process, Progress, and Format) explain and predict the overall item difficulties of the Carbon Cycle items in practice.

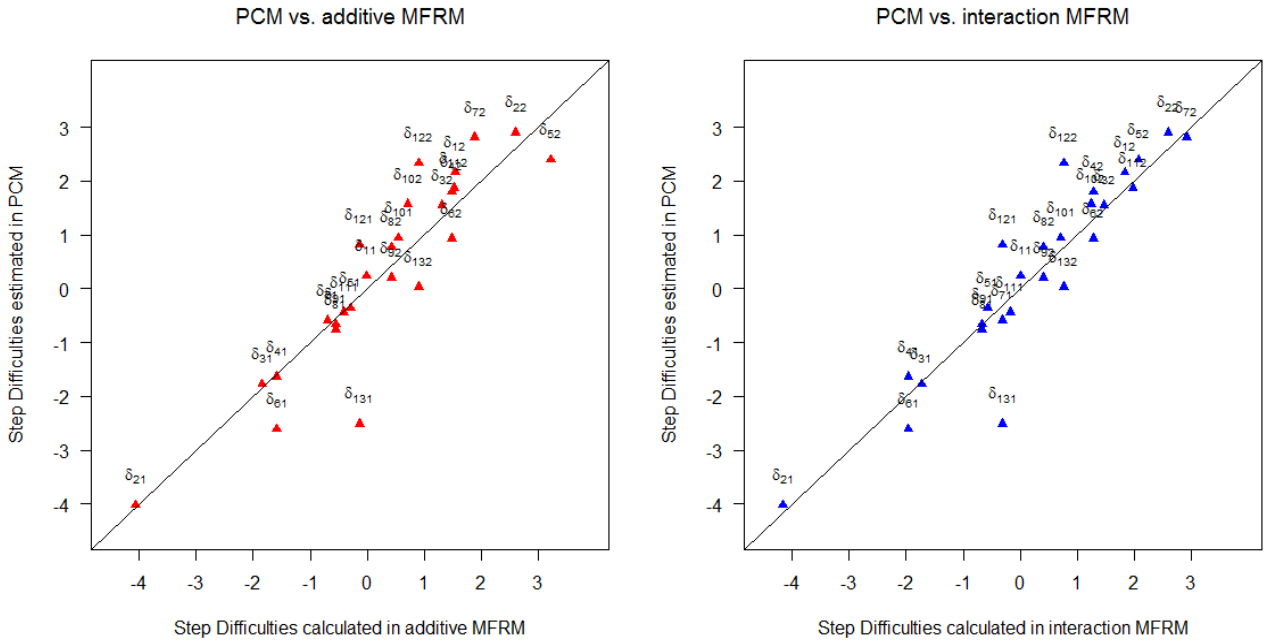


Figure 2-3. Graphical comparison of step difficulties for the Carbon Cycle items in item explanatory models

Table 2-5 represents the results of the item property effects on the Carbon Cycle items and item fit statistics in the interaction MFRM. Although the step deviation parameters were also estimated in the model, they are not the main concern of interpreting the item property effects in the MFRM-based item explanatory models. Focusing on the effects of item properties on the overall item difficulties, the main effects of each item property and the interaction effect between the Process and Progress properties were examined.

In terms of item fit statistics for the item property effect parameters, the results show the mean square (MNSQ) fit statistics and corresponding 95% confidence interval. The MNSQ fit statistic is calculated by taking the ratio of the observed variance and the expected variance (Wilson, 2005), so that it provides a comparison of the expected number of students responding in a category with the observed number of responding in that category (Wu et al., 2007). Under the null hypothesis, the MNSQ fit statistic is 1.0. To evaluate the MNSQ as a measure of effect size, the suggested range which are greater than 0.75 and less than 1.33 can be also considered (Adams & Khoo, 1996; Draney & Wilson, 2008; Wilson, 2005). In the results, for all the item property effect parameters, the confidence interval of the MNSQ fit statistics contains 1.0 and also the MNSQ fit statistics were within the suggested range. Thus, the item fit statistics appeared to be fairly good in the interaction MFRM.

Table 2-5. Item property effects on the Carbon Cycle items and item fit in the interaction MFRM

Predictor	Estimate	SE	Fit statistics (weighted)	
			MNSQ	95% C.I.
Main effect of Process property				
Cellular respiration (CR)	0.445∗	0.042	0.99	(0.91, 1.09)
Photosynthesis (PS)	-0.082∗	0.027	0.91	(0.92, 1.08)
Digestion/biosynthesis (DB)	0.558∗	0.048	0.96	(0.91, 1.09)
Main effect of Progress property				
Large-scale systems (LS)	-0.581∗	0.031	0.99	(0.74, 1.26)
Micro-scale systems (MS)	0.058∗	0.024	0.96	(0.92, 1.08)
Energy (EN)	-0.006	0.025	1.10	(0.92, 1.08)
Mass (MA) †	0.529∗	0.046	1.06	(0.92, 1.08)
Main effect of Format property				
Multiple choice (MC)	0.375∗	0.020	1.00	(0.91, 1.09)
Yes/No choice (YN) †	-0.375∗	0.020	1.06	(0.92, 1.08)
Interaction effect between Process and Progress				
CR × LS	-0.272∗	0.068	0.99	(0.74, 1.26)
CR × MS	0.768∗	0.053	0.97	(0.89, 1.11)
PS × MS	0.267∗	0.032	1.01	(0.92, 1.08)
DB × MS	-0.583∗	0.048	0.89	(0.92, 1.08)
CR × EN	0.105∗	0.053	0.98	(0.89, 1.11)
PS × EN	-0.418∗	0.030	1.05	(0.93, 1.07)
CR × MA †	-0.601∗	0.101	1.09	(0.90, 1.10)
PS × MA †	0.151∗	0.044	1.08	(0.92, 1.08)
DB × MA †	0.583∗	0.048	0.92	(0.89, 1.11)

Note: † parameter estimates are constrained by effect coding. ∗$p < .05$ (significant at the 5% level).

Regarding the item property effects, except for energy in the Progress main effect, all other predictors were statistically significant at the 5% significance level. In terms of investigating the

main effect of the Process property, holding other properties constant, cellular respiration, photosynthesis, and digestion/biosynthesis processes contributed to the overall item difficulties by 0.445, -0.082, and 0.558 logits respectively. This implies that the items questioning the photosynthesis process were the least difficult for students to answer at higher levels of the learning progression for carbon cycling, whereas the items regarding the digestion/biosynthesis process were the most difficult.

For the Progress main effect, keeping other properties constant, large-scale systems (-0.581) contributed the least and mass (0.529) contributed the most to the overall item difficulties, and micro-scale systems (0.058) and energy (-0.006) were between them. The interpretation is that the items questioning large-scale systems were the least difficult and the items regarding mass were the most difficult for students to answer at higher levels of the learning progression for carbon cycling, and the items about micro-scale systems or energy were moderately difficult.

For the Format property, holding other properties constant, multiple choice with explanation format made the overall item difficulty of an item 0.750 logits more difficult than yes/no choice with explanation format, meaning that multiple choice formatted items were more difficult for students to answer at higher levels of the learning progression for carbon cycling than yes/no choice formatted items.

In addition, there was a significant interaction effect on the overall item difficulties between the Process and Progress properties. For example, keeping other properties constant, the cellular respiration process which occurs in micro-scale systems and energy made the overall item difficulty of an item more difficult by 0.768 and 0.105 logits respectively, and the cellular respiration process which occurs in large-scale systems and mass made it less difficult by 0.272 and 0.601 logits respectively.

These results of the item property effects on the Carbon Cycle items were portrayed on a Wright map as in Figure 2-4. The leftmost panel shows the estimated latent ability distribution and the other panels show estimates of the item property effects on the overall item difficulties. This informative map facilitates comparisons of the main or interaction effects within and between the three item properties. In particular, the effects of each item property are interpreted as compared with the students' latent abilities. For instance, the third panel for the Progress main effect indicates that students with the average (zero) latent ability are likely to answer, with more than 50% probability, at higher levels of the learning progression for carbon cycling when the items are questioning large-scale systems. Those students are likely to answer at higher levels with less than 50% probability when the items are questioning mass. When the items are questioning micro-scale systems or energy, those students are likely to answer at higher levels with near 50% probability.

By considering a MFRM-based item explanatory model, the overall item difficulty for a new generalized item is constructed by the estimated item property effects as follows:

$$\beta_i = \alpha_{Process} + \gamma_{Progress} + \omega_{Format} + \varphi_{Process \times Progress} .$$

For example, the overall item difficulty for the item 1 (BODYTEMP) is 0.445 (CR) – 0.006 (EN) + 0.375 (MC) + 0.105 (CR × EN) = 0.919. In addition, the overall item difficulty for newly developing items can be predicted. If one wants to design an item by combining the cellular respiration process, the mass progress variable, and the yes/no choice with explanation format, the overall item difficulty for this item would be 0.445 (CR) + 0.529 (MA) – 0.375 (YN) – 0.601 (CR × MA) = – 0.002.

Logit	Latent Ability	Process	Progress	Format	Process * Progress
3					
	X				
	X				
	X				
2	X				
	XX				
	XX				
	XX				
	XX				
	XXXX				
	XXXX				
	XXXXX				
1	XXXXXXX				
	XXXXX				
	XXXXXXX				CR*MS
	XXXXXX	DB	MA		DB*MA
	XXXXXXXXX	CR			
	XXXXXXXX			MC	PS*MS
	XXXXXXXXX				PS*MA
	XXXXXXXX		MS		CR*EN
0	XXXXXXXXXX	PS	EN		
	XXXXXXX				
	XXXXXXXXX			YN	CR*LS
	XXXXXXXXX				PS*EN
	XXXXXXXXXX		LS		DB*MS
	XXXXXXX				CR*MA
	XXXXXXXX				
-1	XXXX				
	XXXX				
	XXXXX				
	XXX				
	XX				
	XXX				
	XX				
	X				
-2	X				
	X				
	X				
-3					

Each 'X' represents 6.7 cases

Figure 2-4. Wright map for item property effects on the Carbon Cycle items

Consequently, the results of the item property effects enable to check and validate underlying hypotheses applied to the Carbon Cycle assessment construction as well as to develop a new Carbon Cycle item in a scientific and systematic manner. Thus, the MFRM-based item explanatory models can give Carbon Cycle assessment developers empirical evidence for determining construct validity and helpful feedback for enhancing quality of the assessment development.

2.3.3. Application to the Social Evaluative Reasoning data

Data

For the second empirical study, I used a subset of the Social Evaluative Reasoning (SER) in the workplace data (Jolin, 2015, 2018), which were derived from a newly developed instrument designed as a measure of soft skill proficiency for transition-age secondary students with disabilities. The SER in the workplace instrument was developed in a sound and systematic way by using Wilson's (2005) four building blocks approach to constructing measures, which takes advantage of the principles of sound educational measurement. The four building blocks are a cycle of the measurement processes: construct mapping, item design, outcome space, and measurement model.

The higher order SER in the workplace construct is defined as context-specific critical thinking, which involves evaluation of the effectiveness and appropriateness of an employee's behaviors as it occurs in response to common antecedent conditions in entry-level employment heavy in soft skill demand (Jolin, 2018). Within a workplace setting, proficiency in the SER ability is conceptualized in an Antecedent-Behavior-Consequence formulation: some amount of available social information in antecedent events (e.g. a customer's verbal/nonverbal social cues), a target employee's behavioral response, and the appropriateness of that response given the situational context of the workplace and the organizing and directing forces it places upon employee behaviors.

In developing the SER instrument, 16 workplace scenarios in a visual narrative format were employed to depict a target employee responding either correctly or incorrectly to the social cues emoted by customers in various workplace settings of entry level employment heavy in soft skill demand. When designing the workplace scenarios, three design factors were hypothesized to contribute to the construct complexity of the SER items. The first factor is the Emotion which includes four types of emotional cues present in the scenarios—basic, complex, both, none. The second factor, the Language, has three types of language used to deliver social cues—literal, figurative, and none. The third factor, the Resolution, is outcome resolution by the target employee—correct resolution and incorrect resolution.

The construct complexity of the items was manipulated on the basis of the frequency and co-occurrence of these three factors when designing the workplace scenarios. For instance, the 4th workplace scenario described as "Understanding figurative language: Sweet tooth idiom" in Figure 2-5 was designed to depict a target employee responding correctly to a figurative language cue—no Emotion factors were targeted in this scenario. Furthermore, the scenario had a correct resolution (i.e., the employee provided an appropriate response to the customer's instance of figurative language use). Overall, this scenario can be identified as a combination of no emotional

cue in the Emotion factor, figurative language in the Language factor, and correct resolution in the Resolution factor.



Figure 2-5. The 4th workplace scenario for understanding figurative language: sweet tooth idiom. Adapted from *"Social Evaluative Reasoning in the Workplace: Validation of an Assessment of Soft Skill Proficiency for Secondary Students in Special Education,"* by J. Jolin, 2018, unpublished doctoral dissertation. University of California, Berkeley, p. 116. Copyright 2018 by Jerred Jolin.

Each scenario was followed by two item stems associated with two target measurement outcomes: (1) "List all the social cues available to the employee in this scenario", and (2) "Overall, did the employee do the right thing in this scenario? Why?". The first item stem generated 16 items to measure the Social Perceptual Unit (SPU) detection ability, and the second item stem produced 16 items to measure the Evaluative Inference (EI) ability. In this empirical study, the evaluative inferencing items were used to investigate item property effects on the overall item difficulties. In total, 80 participants who are transition-age young adults with disabilities responded to the 16 evaluative inferencing items and their responses were scored as three ordered levels of evaluative inferencing ability based on the inferencing taxonomy by Pearson (1978).

Figure 2-6 presents a construct map for evaluative inferencing ability with corresponding outcome space. This outcome space categorizes the responses to the evaluative inferencing items requiring the evaluation of the appropriateness of a target employee's behavior, so that it is functioning as a scoring rubric for the items. Category level 0 covers responses providing an incorrect evaluation of the resolution of the scenario outcome, or a correct but incomplete or irrelevant evaluation. Category levels 1 and 2 represent increasing levels of evaluative inferencing ability to make accurate inferences about the resolution of the scenario outcomes based on the categories of information that respondents use to support their inferences.

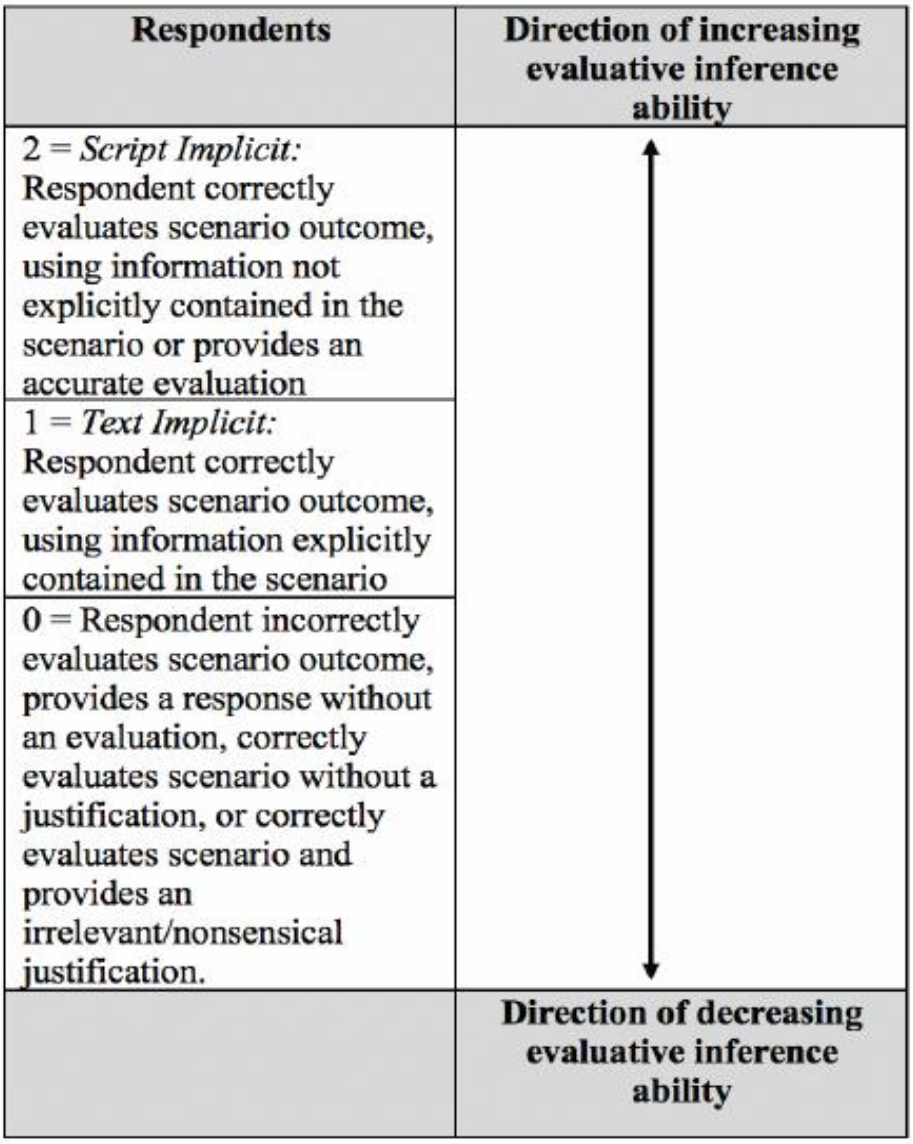


Figure 2-6. Construct map for evaluative inferencing ability with corresponding outcome space. Adapted from "*Social Evaluative Reasoning in the Workplace: Validation of an Assessment of Soft Skill Proficiency for Secondary Students in Special Education,"* by J. Jolin, 2018, unpublished doctoral dissertation. University of California, Berkeley, p. 53. Copyright 2018 by Jerred Jolin.

In particular, the three scenario design factors are regarded as categorical item properties or sub-facets within the item facet in MFRM-based item explanatory models. The 16 evaluative inferencing items based on each scenario are classified by predictors of each item property as in Table 2-6. These predictors are working as weights of the elementary components gathered into a Q matrix in item explanatory models such as the LLTM (see Kubinger, 2009; Poinstingl, 2009). Thus, each item is modeled as a new generalized item constructed by the three categorical item properties in item explanatory models.

Table 2-6. Q matrix for the evaluative inferencing items composed of three item properties

	Emotion				Language			Resolution	
Item	Basic	Complex	Both	None	Literal	Figurative	None	Correct	Incorrect
1	0	0	1	0	1	0	0	0	1
2	0	0	0	1	0	1	0	1	0
3	0	0	1	0	1	0	0	0	1
4	0	0	0	1	0	1	0	1	0
5	0	1	0	0	0	0	1	0	1
6	1	0	0	0	0	1	0	1	0
7	0	0	1	0	0	1	0	0	1
8	0	0	1	0	1	0	0	1	0
9	0	0	0	1	0	1	0	0	1
10	0	0	1	0	1	0	0	1	0
11	1	0	0	0	0	1	0	1	0
12	0	1	0	0	0	1	0	0	1
13	0	1	0	0	0	0	1	1	0
14	1	0	0	0	0	1	0	0	1
15	0	0	1	0	1	0	0	0	1
16	0	0	1	0	1	0	0	0	1

Note. Adapted from "*Social Evaluative Reasoning in the Workplace: Validation of an Assessment of Soft Skill Proficiency for Secondary Students in Special Education*," by J. Jolin, 2018, unpublished doctoral dissertation. University of California, Berkeley, p. 51. Copyright 2018 by Jerred Jolin.

Results

For data analysis, the PCM and two item explanatory models using the MFRM approach were fitted to the SER in the workplace data using the MRCML model framework. In order to identify contributions of the three design factors within the 16 scenarios to the overall item difficulties of the evaluative inferencing items, I compared a MFRM with additive item property effects (called the additive MFRM here) to a MFRM with additive item property effects and two interactions between the Emotion and Resolution properties and between the Language and Resolution properties (called the interaction MFRM here). Considering the outcome space for evaluative inferencing ability, the scale structure was determined by the three item properties in the two item explanatory models.

Table 2-7 presents the results of goodness-of-fit between the fitted models to see performance of the two MFRM-based item explanatory models. The LR test comparing with the PCM was significant for the additive MFRM ($\chi^2(15) = 36.81$, $p = .001$) as well as for the interaction MFRM ($\chi^2(9) = 31.16$, $p < .001$), which represents that both item explanatory models fit worse than the PCM as expected. This result was confirmed by the AIC in that its values for the two item explanatory models were greater than the PCM (1260.70 for the PCM, 1267.51 for the additive MFRM, 1273.86 for the interaction MFRM). This makes sense because a smaller number

of estimated item parameters in item explanatory models made model fit worse. 33 parameters estimated in the PCM were reduced to 18 in the additive MFRM and 24 in the interaction MFRM.

Table 2-7. Goodness-of-fit of item explanatory models to the Social Evaluative Reasoning data

Models		PCM	Item explanatory models	
			additive MFRM	interaction* MFRM
Good-ness of Fit	Deviance	1194.70	1231.51	1225.86
	AIC	1260.70	1267.51	1273.86
	BIC	1339.30	1310.39	1331.02
	k	33	18	24

Note: k = number of estimated parameters; * interactions between the Emotion and Resolution properties and between the Language and Resolution properties.

However, this result was not supported by the BIC because its values for the PCM was greater than the two item explanatory models (1339.30 for the PCM, 1310.39 for the additive MFRM, 1331.02 for the interaction MFRM). This conflicted result could be figured out reasonably by considering the difference between the AIC and BIC, although it doesn't look as expected. The BIC favors parsimonious models to a greater extent than the AIC does. While the BIC penalizes complexity more heavily for the number of freely estimated parameters, the AIC penalizes the larger model much less than the BIC (Kuha, 2004). Basically, they differ in the meaning of penalty terms, assumptions, objectives, and theoretical motivations (Kuha, 2004; Wagenmakers & Farrell, 2004). In terms of the BIC, we could say that the two MFRM-based item explanatory models performed better than the PCM in this empirical study. This result implies that the item design with the three design factors worked well for item generation so that they made parsimonious fit to the data without severe loss of model fit (accuracy).

When comparing between the two item explanatory models, the LR test was not significant as $\chi^2(6) = 5.65$ ($p = 0.463$), meaning that the interaction effects of the Resolution with the Emotion or the Language properties were not substantial. Also, both the AIC and BIC indices indicate that the interaction MFRM fitted worse than the additive MFRM.

In addition to the statistical goodness-of-fit comparison, effect size was also investigated: a graphical comparison and a correlation analysis were conducted to examine agreement for the estimated and calculated step difficulties between the fitted models. Table 2-8 presents the step difficulties for the evaluative inferencing items. They were freely estimated in the PCM, whereas they were calculated by the estimated item parameters in the two MFRM-based item explanatory models.

In Figure 2-7, the step difficulties for the original items in the PCM and corresponding values for the new generalized items in each item explanatory model were compared graphically. In this graphical comparison, although the interaction MFRM had a few more step difficulty points which were located close to the 45-degree line indicating perfect alignment than the additive MFRM, their agreement performances appeared to be comparable. In both models, the first step difficulty of item 3 was located farthest from the 45-degree line. Such comparable performance was also

found from a correlation analysis. Correlations of step difficulties with the PCM were similar in the two item explanatory models (ρ = 0.74 for the additive MFRM, and ρ = 0.76 for the interaction MFRM). In addition, the proportion of explained step difficulty variance (R^2) was 0.53 for the additive MFRM and 0.55 for the interaction MFRM. This indicates that the interaction MFRM accounted for variability in the step difficulties of the PCM more by 2% than the additive MFRM.

Table 2-8. Step difficulties for the evaluative inferencing items in item explanatory models

Item	Step difficulty	PCM	additive MFRM	interaction MFRM
1	δ_{11}	-0.54	0.66	0.77
	δ_{12}	0.15	-0.05	0.12
2	δ_{21}	-0.17	-0.18	-0.15
	δ_{22}	0.01	-0.04	-0.01
3	δ_{31}	2.50	0.66	0.77
	δ_{32}	-0.25	-0.05	0.12
4	δ_{41}	-0.15	-0.18	-0.15
	δ_{42}	0.04	-0.04	-0.01
5	δ_{51}	-0.04	0.05	-0.03
	δ_{52}	-0.39	-0.38	-0.42
6	δ_{61}	1.60	0.55	0.63
	δ_{62}	-0.61	0.03	0.16
7	δ_{71}	0.22	0.22	0.22
	δ_{72}	0.07	0.03	0.05
8	δ_{81}	-0.25	0.08	-0.07
	δ_{82}	0.39	0.64	0.47
9	δ_{91}	0.13	0.18	0.15
	δ_{92}	-0.52	-0.52	-0.53
10	δ_{101}	0.10	0.08	-0.07
	δ_{102}	0.61	0.64	0.47
11	δ_{111}	-0.11	0.55	0.63
	δ_{112}	0.71	0.03	0.16
12	δ_{121}	0.53	0.57	0.55
	δ_{122}	-0.83	-0.83	-0.84
13	δ_{131}	-1.17	-1.30	-1.14
	δ_{132}	1.27	1.08	1.26
14	δ_{141}	0.87	1.30	0.90
	δ_{142}	-1.23	-0.84	-1.24
15	δ_{151}	0.73	0.66	0.77
	δ_{152}	0.56	-0.05	0.12
16	δ_{161}	0.66	0.66	0.77
	δ_{162}	-0.20	-0.05	0.12

Putting the results of model comparison together, the additive MFRM performed better than the interaction MFRM in the SER in the workplace data. Despite the comparable performance in terms of the agreement of parameter estimates and the explained variability in the step difficulties,

the additive MFRM had a superior goodness-of-fit to the interaction MFRM. Consequently, the additive MFRM was identified as the final item explanatory model to investigate further how the three item properties (Emotion, Language, and Resolution) explain and predict the overall item difficulties of the evaluative inferencing items in the SER instrument.

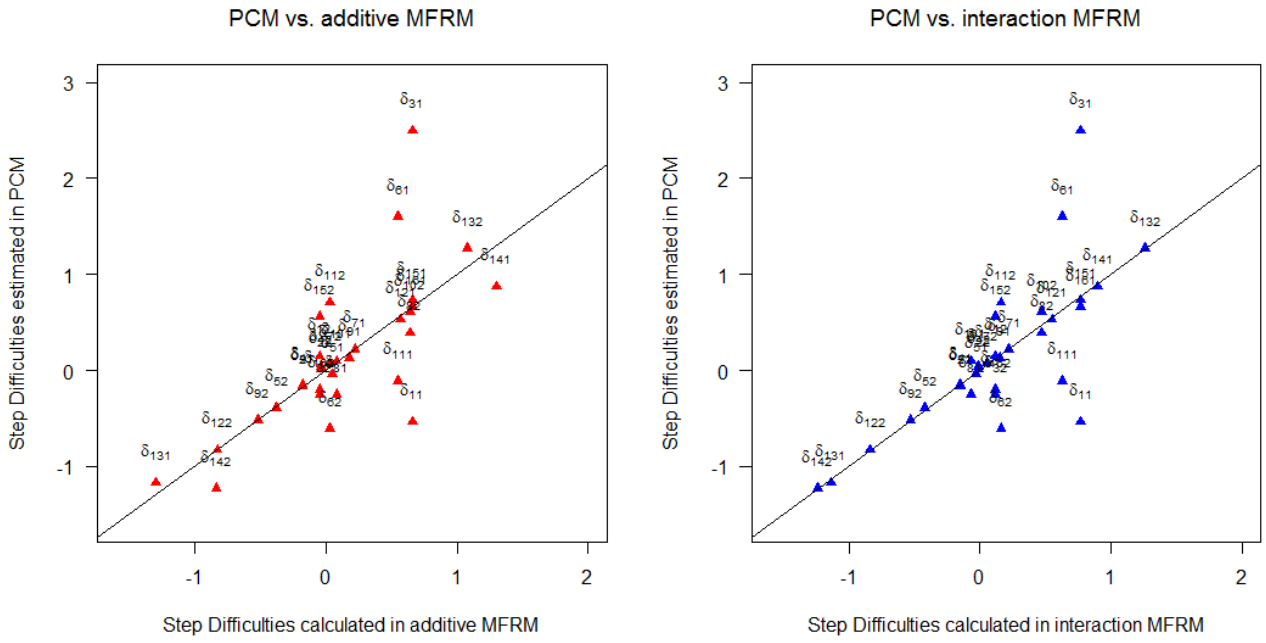


Figure 2-7. Graphical comparison of step difficulties for the evaluative inferencing items in item explanatory models

Table 2-9 presents the results of the item property effects on the evaluative inferencing items and item fit statistics in the additive MFRM. In item explanatory models using the MFRM approach, the step deviation parameters are not the main concern of interpreting the item property effects although they were also estimated in the model. The main effects of each item property on the overall item difficulties were examined in the additive MFRM.

As examining item fit statistics for the item property effect parameters, the results show the mean square (MNSQ) fit statistics and corresponding 95% confidence intervals. For all the item property effect parameters, the MNSQ fit statistics appeared to be within the suggested range, which are greater than 0.75 and less than 1.33 (Adams & Khoo, 1996; Draney & Wilson, 2008; Wilson, 2005). Also, all the confidence intervals of the MNSQ fit statistics contained 1.0 which is under the null hypothesis. Thus, the item fit statistics appeared to be fairly good in the additive MFRM.

In investigating the item property effects, only the "Basic" and "Both" emotional cues in the Emotion main effect were statistically significant at the 5% significance level. Note that standard errors of parameter estimates are affected by the sample size in statistical models. The small sample size of 80 could be one of the reasons for such low rate of significant predictors in this empirical study. Nevertheless, the estimated effects of each item property on the overall item difficulties were interpreted to see how they works in practice.

Regarding the main effect of the Emotion property, holding other properties constant, Basic and Both emotional cues made the overall item difficulty of an item more difficult by 0.305 and 0.203 logits respectively, and Complex and None emotional cues made it less difficult by 0.054 and 0.095 logits respectively. This means that the items depicting workplace scenarios in which the Basic or Both emotional cues present were more difficult for respondents to answer at higher levels of evaluative inferencing ability, and the items targeting evaluative inferencing in the workplace scenario designed with Complex or None emotional cues were less difficult.

Table 2-9. Item property effects on the evaluative inferencing items and item fit in the additive MFRM

Predictors	Estimate	SE	Fit statistics (weighted)	
			MNSQ	95% C.I.
Main effect of Emotion property				
Basic	0.305∗	0.117	1.22	(0.71, 1.29)
Complex	-0.054	0.124	1.01	(0.70, 1.30)
Both	0.203∗	0.085	1.16	(0.69, 1.31)
None	-0.095	0.133	1.05	(0.71, 1.29)
Main effect of Language property				
Literal	0.128	0.083	1.01	(0.70, 1.30)
Figurative	-0.046	0.067	1.04	(0.69, 1.31)
None †	-0.083	0.106	0.87	(0.74, 1.26)
Main effect of Resolution property				
Correct	0.030	0.055	0.95	(0.69, 1.31)
Incorrect †	-0.030	0.055	1.09	(0.69, 1.31)

Note: † parameter estimates are constrained by effect coding. ∗p < .05 (significant at the 5% level).

For the Language main effect, keeping other properties constant, Literal, Figurative, and None languages contributed to the overall item difficulties by 0.128, -0.046, and -0.083 logits respectively. This implies that the items in which Literal language is used to deliver social cues were more difficult for respondents to answer at higher levels of evaluative inferencing ability, and that the items targeting evaluative inferencing in the workplace scenario designed with Figurative or None languages for social cue delivery were less difficult.

Regarding the Resolution property, holding other properties constant, Correct resolution made the overall item difficulty of an item 0.060 logits more difficult than Incorrect resolution. This means that the items depicting workplace scenarios designed with Correct outcome resolution by the target employee were more difficult for respondents to answer at higher levels of evaluative inferencing ability than those designed with Incorrect outcome resolution by the target employee.

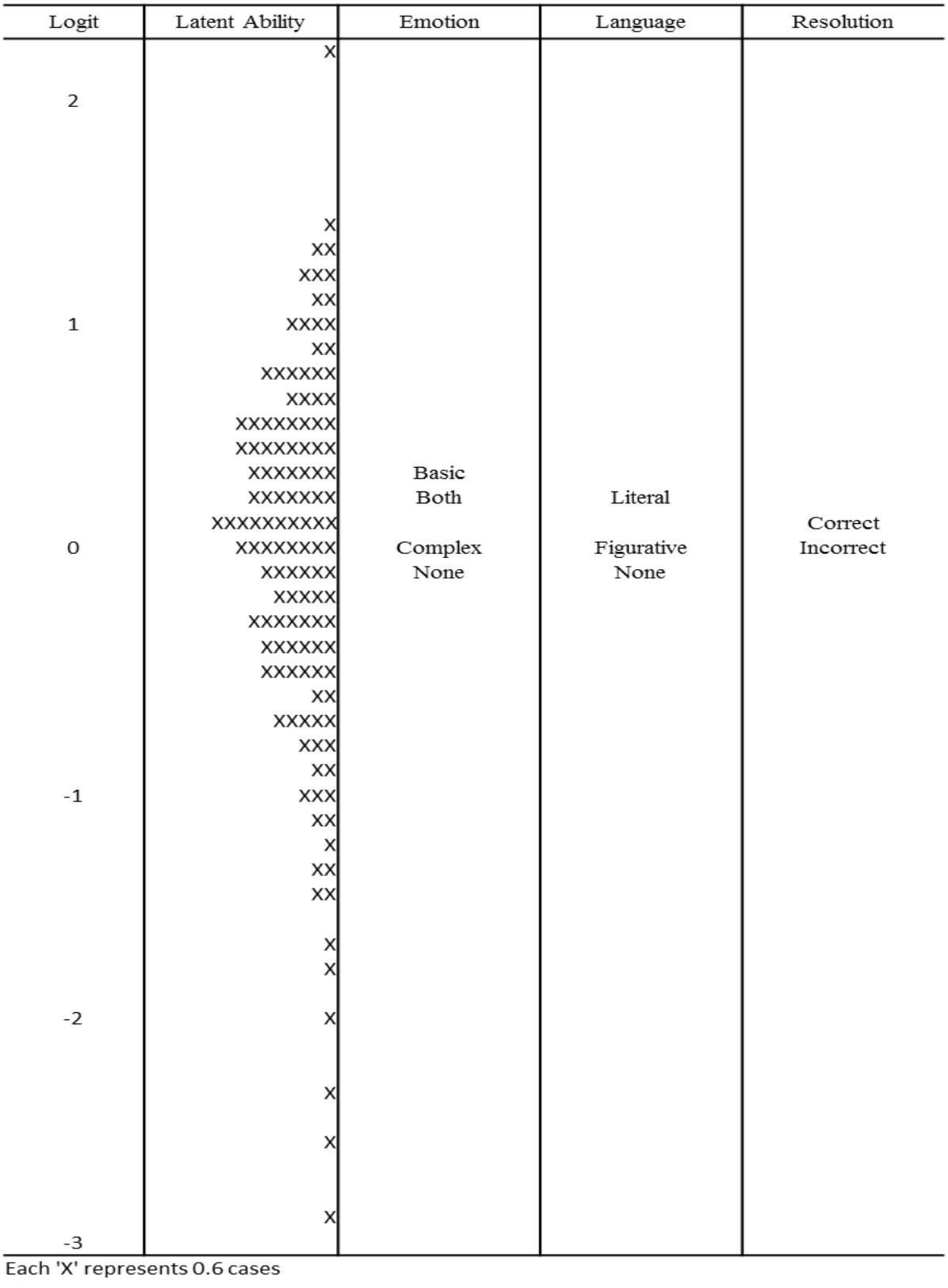


Figure 2-8. Wright map for item property effects on the evaluative inferencing items

In order to facilitate comparisons of the main effects within and between the three item properties, these results of the item property effects on the evaluative inferencing items were displayed on a Wright map as in Figure 2-8. The estimated latent ability distribution was portrayed in the leftmost panel, and the estimated main effects of each item property on the overall item difficulties were displayed in each of the other panels. As compared with the respondents' latent abilities, the item property effects are interpreted. For instance, the third panel for the Language main effect indicates that respondents with the average (zero) latent ability are likely to answer, with more than 50% probability, at higher levels of evaluative inferences about the resolution of the scenario outcomes when the items are designed with None language. Those respondents are likely to answer at higher levels with less than 50% probability when the items are designed with Literal language. When the items are designed with Figurative language, those respondents are likely to answer at higher levels with near 50% probability.

Lastly, with consideration for a MFRM-based item explanatory model with no interactions, the overall item difficulty for a new generalized item is constructed by adding the estimated main effects of the three item properties as follows:

$$\beta_i = \alpha_{Emotion} + \gamma_{Language} + \omega_{Resolution} .$$

For example, the overall item difficulty for the item 4 based on the 4th workplace scenario in Figure 2-5 (Understanding figurative language: Sweet tooth idiom) is – 0.095 (None emotional cue) – 0.046 (Figurative language) + 0.030 (Correct resolution) = – 0.111. In addition, the overall item difficulty for newly developing items can be predicted. If one wants to design an item by combining Basic emotional cue, Literal language, and Correct outcome resolution, the overall item difficulty for this item would be 0.305 (Basic emotional cue) + 0.128 (Literal language) + 0.030 (Correct resolution) = 0.463.

Consequently, these results enable to develop a new evaluative inferencing item for the SER in the workplace instrument in a scientific and systematic manner. Also, these results enable to check and validate underlying hypotheses applied to the evaluative inferencing ability construct. Jolin (2018) demonstrated empirical validity evidences for the internal structure of the evaluative inferencing items using the MFRM-based item explanatory models which can provide informative feedback for enhancing quality of the instrument development.

2.4. Conclusion and Discussion

I have investigated how to apply item explanatory models using the MFRM approach to polytomous data under the MRCML model framework, given the ordered-category responses and adjacent-categories logits. In the two empirical studies, applications of the MFRM-based item explanatory models to the Carbon Cycle assessment data and the Social Evaluative Reasoning data were demonstrated respectively.

In terms of the goodness-of-fit, the agreement of parameter estimates, and the explained variability in the step difficulties, the two MFRM-based item explanatory models performed differently between the two empirical studies. The first empirical study demonstrated that the interaction MFRM performed better than the additive MFRM on the Carbon Cycle items, and the

two item explanatory models fit the data worse than the PCM as expected. In contrast, the second empirical study showed that the additive MFRM performed better than the interaction MFRM on the evaluative inferencing items, and compared to the PCM, the two item explanatory models had an inferior goodness-of-fit in terms of the AIC but a superior goodness-of-fit in terms of the BIC. This conflicted result between the two goodness-of-fit indices hints that the item design with the three design factors worked well for item generation because the BIC indicates parsimonious fit to the data without severe loss of model fit (accuracy).

Based on the performance of model comparison, the final item explanatory model to investigate the item property effects was differently identified between the two studies. In the first study, an interaction effect between the Process and Progress properties appeared to be significant on the Carbon Cycle items and hence the interaction MFRM was chosen. In the second study, two interaction effects of the Resolution with the Emotion and the Language properties were not substantial on the evaluative inferencing items and hence the additive MFRM was chosen. According to the different final models, we could see how the observed item properties explain and predict the overall item difficulties differently in practice between the two studies.

Item explanatory models can enhance explanatory inferences from various types of item responses. The MFRM-based item explanatory models can provide explanatory inferences from the ordered-category responses. Interestingly, although the number of the original items were different between the two empirical data sets (13 Carbon Cycle items and 16 evaluative inferencing items), based on the same number of predictors of the observed item properties, the same number of item property effect parameters could be estimated in the MFRM-based item explanatory models (but the number of step deviation parameters can vary according to the scale structure). This represents a methodological advantage of item explanatory models in extracting essential and meaningful elementary components from the item response data. By incorporating item properties, these models enable us to estimate a smaller number of item parameters than the saturated model to explain and predict the item effects, which has an explanatory value (De Boeck & Wilson, 2004; Kubinger, 2009).

Furthermore, item explanatory models can give us informative and practical feedback for enhancing quality of developing items in an instrument, test, and assessment in educational and psychological measurement practices. In item development and substantive interpretation aspects, they are helpful to figure out what property of the items such as a cognitive operation and learning content makes it difficult for students to answer them more correctly or achieve intended learning goals. In the two empirical studies, the effects of item properties on the overall item difficulties were investigated as shown on the Wright maps in Figure 2-4 and Figure 2-8, so that we could figure out how the item properties works in relation to item design or construction. This suggests that the MFRM-based item explanatory models can be used to validate a rationale applied to the item design or construction as well as to provide empirical evidence for determining construct validity (e.g. validity evidence for the internal structure) in the context of a particular assessment (Embretson & Gorin, 2001; Jolin, 2018; Poinstingl, 2009). Moreover, they can allow us to develop a new item in a more scientific and systematic manner rather than an intuitive manner. By constructing a combination of the estimated item property effects in item generation process, we can predict the overall difficulty of the newly developed items.

Even though I have discussed various advantages and implications of item explanatory models using the MFRM approach, there are some additional measurement issues that can be addressed in future studies. In particular, the MFRM-based item explanatory models used in this

research are essentially fixed item effects IRT models in that the item location parameters are decomposed into linear combinations of the fixed item property effects. However, they have an unrealistic assumption that the overall item difficulties are completely explained by item properties. Future studies need to address the value of random-effects versions extended with consideration of random item variations (Mislevy, 1988; Janssen, Schepers, & Peres, 2004), which will allow for increasing the precision of estimation of the item parameters. In addition, although I assumed a unidimensional model, multidimensional extensions of item explanatory models such as the Random-Weights Linear Logistic Test Model (RW-LLTM; Rijmen& De Boeck, 2002) can be considered in order to examine interactions between item properties and persons.

Chapter 3.
Polytomous Item Explanatory IRT Models

3.1. Introduction

In Item Response Theory (IRT), both person traits and item characteristics are modeled together as covariates in a statistical model with a nonlinear relationship between the item response probability and the covariates. For example, the Rasch model imposes a probability of success for an item with an inverse logistic transformation of the difference between person ability and item difficulty parameters. This model structure hints that one can explain either side of the item response data, namely person side or item side, or both sides by incorporating person properties, item properties, and their interactions as covariates in the statistical model. Explanatory item response models (EIRM; De Boeck & Wilson, 2004) address various types of explanatory IRT models under the Generalized Linear Mixed Modeling (GLMM; Breslow & Clayton, 1993; De Boeck & Wilson, 2004; McCulloch & Searle, 2001) framework. Among the EIRM approaches, item explanatory IRT models focus on explaining the item side of the item response data in which item properties are incorporated to account for the item effects. To avoid losing research focus, I leave it to future studies to investigate further person explanatory or doubly explanatory models and their multidimensional extensions as well as two-parameter or three-parameter IRT models.

A typical case of item explanatory IRT models is the Linear Logistic Test Model (LLTM; Fischer, 1973). This approach is useful for explaining item difficulties by incorporating item properties such as item design variables, cognitive operations, item content features, and response formats. In other words, by decomposing the item difficulty parameters of the Rasch model into a linear combination of elementary components related to the item properties (Embretson & Reise, 2000), it can help predict success on a particular type of item. The LLTM can also be used as a testing tool for hypothesized constructs in item generation. For example, when a specific test construction rationale which is hypothesized for item generation is developed to predict item difficulties, the appropriateness of that rationale can be assessed by means of the LLTM so that it can help enable automatized item generation (Reif, 2012). Moreover, as relevant research using this approach increases, other potential applications have been developed and studied theoretically and empirically. That is, the LLTM approach can be used effectively for psychometric studies on measuring the effect of various testing conditions such as item presentation position, content-specific learning, speeded item presentation, and item response format as well as the measurement of changes (Cristante & Robusto, 2007; Kubinger, 2009). Thus, item explanatory models can serve to provide informative feedback for enhancing test development, item generation, and cognitive assessment.

Despite such diverse potential uses of the LLTM approach in educational measurement research, it has been less commonly used. Historically, most of the early studies using the LLTM often appeared in German-speaking countries and were written in German (Kubinger, 2009; Lei, 2009; Sonnleitner, 2008). Furthermore, although the observed item responses are typically categorical, either dichotomous or polytomous, most of its applications are actually limited to dichotomous (or binary) item responses (e.g., De Boeck, 2008; Fischer, 1973; Kubinger, 2009;

Poinstingl, 2009). However, it is very common to have polytomous (or multi-categorical) data such as ordered-category responses or nominal responses in a wide range of educational, psychological, and sociological applications (De Boeck & Wilson, 2004). In particular, I will focus on ordered-category (or ordinal) responses for polytomous data because they are quite common achievement outcomes in educational assessment contexts. For instance, very often the scoring rubrics under a learning progression framework (LPF) will have ordered categorical levels showing qualitatively different learning progression levels, which result in ordered-category responses (e.g. Jin, Shin, Johnson, Kim, & Anderson, 2015). Also, I will focus on the adjacent-categories logit relationship between the ordered-category item responses and covariates of person ability and item characteristics. Adjacent-categories logit based item response models can make person and item parameters separable, i.e. parameter separability, and hence permit specifically objective comparisons of persons and items, i.e. specific objectivity (Masters, 1982). They are appropriate where local comparison between ordered adjacent categories is of interest (Masters & Wright, 1997). Moreover, adjacent-categories logit ordinal regression models have a suitable parameterization for subjectively assessed scores in ordered responses (Anderson, 1984; Johnson, 2007).

My research aim is to investigate how the LLTM approach can be applied to polytomous data, particularly in terms of item explanatory extensions of the polytomous Rasch models under the general statistical modeling framework, which are also labelled as polytomous item explanatory IRT models. One may say that any item explanatory IRT model applied to dichotomous data could be simply applied to polytomous data. However, this is not the case conceptually and practically. In particular, applying the LLTM approach to polytomous data is not that easy due to the complication of item parameterization in the statistical model. The seminal study by Fischer (1977) and its follow-up studies by Fischer and Parzer (1991) and Fischer and Ponocny (1994) investigated item explanatory extensions of polytomous item response models using the LLTM approach. They utilized a normalization constant and basic parameter for item parameterization, however, these parameters are difficult to interpret and even make it hard to calculate and interpret the step difficulties. Masters (1982) pointed out that "To my knowledge, this polytomous generalization of the linear logistic test model (Fischer, 1972) has not been applied." In addition to the complication, the difficulty in re-parameterization incorporating item properties may lead applications of the LLTM approach to polytomous data to be considerably less common than to dichotomous data.

In order to achieve the aim of this research, unidimensional polytomous item response models will be reviewed as a starting point through the following sections. Next, I will investigate how to develop polytomous item explanatory IRT models using the LLTM approach in terms of different item parameterizations, given ordered-category responses and the adjacent-categories logit link. And then, two empirical studies that focus on applying these models to the Carbon Cycle assessment data and to the verbal aggression data, respectively, will be demonstrated. In the empirical studies, we can see how the proposed polytomous item explanatory models work in practice and how item properties explain and predict the difficulties of polytomous items.

3.2. Polytomous Extensions of the Rasch Model under the Multivariate Generalized Linear Mixed Modeling Framework

3.2.1. The Multivariate Generalized Linear Mixed Modeling Framework

Even though basic IRT models for polytomous data under the Multivariate Generalized Linear Mixed Modeling (MGLMM; Agresti, 2002; Hartzel, Agresti, & Caffo, 2001) framework were addressed in De Boeck and Wilson (2004), item explanatory extensions of specific polytomous item response models were not fully discussed. In polytomous data, response categories are either ordinal or nominal. Commonly, to link between probabilities of the observed item responses and covariates for both person and item sides of the data, baseline-category logits are used for nominal responses, and cumulative logits or adjacent-categories logits are used for ordinal responses under the MGLMM framework (De Boeck & Wilson, 2004; Hartzel, Agresti, & Caffo, 2001; Mellenbergh, 1995).

Since ordered-category (or ordinal) item responses are a major interest as mentioned in the introduction section, cumulative logits or adjacent-categories logits will be considered. In the IRT models based on cumulative logits, Samejima's Graded Response Model (GRM; Samejima, 1969) is well-known. Although the cumulative logit link is relatively more popular and widely used among ordinal regression models in statistics and biostatistics (Johnson, 2007), adjacent-categories logit link is more useful when contrasting probabilities of the responses in pairs of adjacent categories (Hartzel, Agresti, & Caffo, 2001). Particularly in educational measurement and psychometrics, ordinal item response models based on adjacent-categories logits are more commonly used, because of the parameter separability by sufficient statistics and hence the possibility of specifically objective comparisons of persons and items (Masters, 1982; Masters & Wright, 1997) as well as the suitability of parameterization for subjectively assigned scores in ordered responses (Anderson, 1984; Johnson, 2007), the local comparison between ordered adjacent categories (Masters & Wright, 1997), the academic tradition (De Boeck & Wilson, 2004), and also the richness of available IRT software packages.

Under the MGLMM framework, there are three components of the statistical model: (a) the distribution of the data, (b) the link function, and (c) the linear predictor. Based on the ordered-category responses, I will exploit the multivariate Bernoulli distribution, namely the multinomial distribution where the total sum of probabilities across response categories equals one. Also, based on the adjacent-categories logits for ordinal item response models, I will employ the adjacent-categories logit link function. For polytomous item explanatory IRT models, the linear predictor is a crucial component to explain item characteristics by incorporating plausible explanatory item properties. I will focus on the linear predictor to apply the LLTM approach to polytomous item response models. The linear predictor per person p in matrix form is

$$\boldsymbol{\eta}_p = \theta_p \boldsymbol{Z}_p - \boldsymbol{X}_p \boldsymbol{\beta}\,, \qquad p = 1, \ldots, P\,, \tag{3.1}$$

where $\boldsymbol{\eta}_p$ is a vector of linear predictors for person p, θ_p is a latent ability of person p, and $\boldsymbol{Z}_p$ is a scoring matrix. $\boldsymbol{X}_p$ is a design matrix, $\boldsymbol{\beta}$ is a corresponding vector of the item parameters, and the negative of their product is considered because it is a tradition for interpreting the item parameters in item response models. The MGLMM is a special case of the Generalized Linear Latent and

Mixed Modeling (GLLAMM; Rabe-Hesketh, Skrondal, & Pickles 2004) framework when a slope or discrimination parameter vector $\boldsymbol{\lambda}$ is set to one (see Zheng & Rabe-Hesketh, 2007). Since I assume unidimensional item response models with no person properties, the scoring matrix $\boldsymbol{Z}_p$ is just a vector of elements of category scores for each item so that it is not a matter of concern in this paper. Instead, the design matrix $\boldsymbol{X}_p$ and the corresponding item parameter vector $\boldsymbol{\beta}$ are important to investigate item explanatory models by incorporating item properties using different item parameterizations.

3.2.2. Polytomous Rasch Models

Based on ordered-category responses and adjacent-categories logits, I will investigate item explanatory extensions of the polytomous Rasch models under the MGLMM framework. To reach those extensions, I will start with a basic model for the adjacent-categories logit based item response models. The Partial Credit Model (PCM; Masters, 1982) is a polytomous extension of the Rasch model, in which the conditional probability that person p with ability θ_p would respond with category score m on item i with step difficulty parameters δ_{im} is defined as:

$$Pr\left(y_{pi} = m \middle| \theta_p\right) = \frac{\exp\left(\sum_{l=0}^{m}(\theta_p - \delta_{il})\right)}{\sum_{h=0}^{M_i}\exp\left(\sum_{l=0}^{h}(\theta_p - \delta_{il})\right)}\,, \qquad m = 0, 1, \ldots, M_i\,, \tag{3.2}$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$, $\delta_{i0} = 0$, and $\sum_{l=0}^{0}(\theta_p - \delta_{il}) = 0$ so that $\sum_{l=0}^{h}(\theta_p - \delta_{il}) = \sum_{l=1}^{h}(\theta_p - \delta_{il})$. Note that for all statistical models in this paper, the mean of person abilities is constrained to zero for model identification[4].

For explanatory convenience and intuitive interpretation of item parameters, the PCM can be described in terms of local comparison in the response probabilities from the adjacent category score $m-1$ to m, which is regarded as the m-th adjacent-categories logit (Tuerlinckx & Wang, 2004). Based on the local comparison, the linear predictor element of the m-th adjacent-categories logit on item i per person p, η_{pim} is expressed as follows:

$$\eta_{pim} = ln\frac{Pr\left(y_{pi} = m \middle| \theta_p\right)}{Pr\left(y_{pi} = m-1 \middle| \theta_p\right)} = \theta_p - \delta_{im}\,, \tag{3.3}$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$, $\delta_{i0} = 0$, and $\eta_{pi0} = 0$. From this local comparison perspective of the PCM, the item parameter δ_{im} is interpreted as a "step" difficulty for scoring m rather than $m-1$ on item i ($i = 1,2, \ldots, \mathrm{I}$). Since a value of the category score can be regarded as a step of switching over to the next response category, the step difficulty parameter δ_{im} indicates the relative difficulty of the m-th step within item i when shifting the category score from $m-1$ to m (Embretson & Reise, 2000; Masters, 1982; Zheng & Rabe-Hesketh, 2007). This is a special case of the multinomial logit model (De Boeck & Wilson, 2004; Glas & Verhelst, 1989).

[4] For model identification, a particular item parameter or the mean of item parameters can be constrained to zero, instead of constraining the mean of person abilities (De Boeck & Wilson, 2004).

In the adjacent-categories logit based item response models, there are three types of predictors that can be included in the design matrix $\boldsymbol{X}_p$ (Tuerlinckx & Wang, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003): (a) item predictors, (b) step predictors[5], and (c) item-by-step predictors.

The item predictor is a predictor if and only if the elements of the corresponding column of the design matrix $\boldsymbol{X}_p$ vary across items but they are fixed as category scores for each item for all persons, the step predictor is a predictor if and only if the elements of the corresponding column of the design matrix $\boldsymbol{X}_p$ vary across steps but they are constant across items for all persons, and the item-by-step predictor is a predictor if and only if the elements of the corresponding column of the design matrix $\boldsymbol{X}_p$ vary across all items and steps but they are constant across persons. Simply put, predictors in the design matrix $\boldsymbol{X}_p$ could be an indicator or an explanatory property of the item specific characteristic (e.g., item location), step specific characteristic (e.g., scale threshold), and item-by-step specific characteristic (e.g., step difficulty), which result in item predictors, step predictors, and item-by-step predictors respectively. For item descriptive models such as the PCM, the item-by-step predictors are all indicators of the step difficulties. For item explanatory models, the predictor matrix will contain explanatory properties for specific item characteristics.

To make a simple illustration of linear predictor vector $\boldsymbol{\eta}_p$ for the PCM under the MGLMM framework, suppose that there are two ordinal items with three response categories each ($M_1 = M_2 = 2$). The scoring matrix $\boldsymbol{Z}_p$ contains all category scores for each item ($m = 0,\ 1,\ 2$). The design matrix $\boldsymbol{X}_p$ contains indicators for the step difficulties, regarded as the item-by-step predictors, and the corresponding item parameter vector $\boldsymbol{\beta}$ contains step difficulty parameters $(\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22})$ which represent item-by-step interaction effects (Tuerlinckx & Wang, 2004). The PCM has the following linear predictor vector for person *p* corresponding to Equation 3.2 and 3.3:

$$\begin{matrix} \boldsymbol{\eta}_p \\ \\ \begin{bmatrix} \eta_{p10} \\ \eta_{p10} + \eta_{p11} \\ \eta_{p10} + \eta_{p11} + \eta_{p12} \\ \eta_{p20} \\ \eta_{p20} + \eta_{p21} \\ \eta_{p20} + \eta_{p21} + \eta_{p22} \end{bmatrix} \end{matrix} = \begin{matrix} = \theta_p \times \boldsymbol{Z}_p - \boldsymbol{X}_p \times \boldsymbol{\beta} \\ \\ \theta_p \times \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \delta_{11} \\ \delta_{12} \\ \delta_{21} \\ \delta_{22} \end{bmatrix}. \end{matrix} \tag{3.4}$$

Note that the linear predictor vector $\boldsymbol{\eta}_p$ contains a cumulative sum of the linear predictor elements of the adjacent-categories logits for all category scores in each item, as shown in the numerator of Equation 3.2. Since each response probability is a function of all linear predictor elements in the denominator, the data matrices must be expanded to have all category scores from zero so that there are $M_i + 1$ rows for each item (Zheng & Rabe-Hesketh, 2007). For example, for the first item,

[5] Tuerlinckx and Wang (2004) used the term "logit predictor", and Rijmen and others (2003) used "category covariates". Instead, we will use the term "step predictor", because a step is more intuitive and pertinent in terms of local comparison approach. See the partial credit model described using the local comparison expression.

when $m = 0$, $\eta_{p10} = 0$,
when $m = 1$, $\eta_{p10} + \eta_{p11} = 0 + (\theta_p - \delta_{11}) = \theta_p - \delta_{11}$, and
when $m = 2$, $\eta_{p10} + \eta_{p11} + \eta_{p12} = 0 + (\theta_p - \delta_{11}) + (\theta_p - \delta_{12}) = 2\theta_p - \delta_{11} - \delta_{12}$.

Interestingly, using a twofold item parameterization, we can split the step difficulty parameter δ_{im} into two item parameters of the item location (a.k.a. overall item difficulty) parameter β_i and the step deviation parameter τ_{im} as follows:

$$\delta_{im} = \beta_i + \tau_{im}, \tag{3.5}$$

so that

$$\eta_{pim} = ln\frac{Pr(y_{pi} = m|\theta_p)}{Pr(y_{pi} = m-1|\theta_p)} = \theta_p - \beta_i - \tau_{im}\,, \qquad m = 1, \dots, M_i\,, \tag{3.6}$$

where $\theta_p \sim N(0, \sigma_\theta^2)$, $\eta_{pi0} = 0$, $\tau_{i0} = 0$, and $\sum_{m=1}^{M_i} \tau_{im} = 0$ so that $\frac{1}{M_i}\sum_{m=1}^{M_i} \delta_{im} = \beta_i$. Here, β_i is an item location parameter for item i, and τ_{im} is a step deviation parameter for the m-th step within item i. Due to the model constraint that $\frac{1}{M_i}\sum_{m=1}^{M_i} \delta_{im} = \beta_i$, the item location parameter is interpreted as the overall item difficulty for each polytomous item. The step deviation parameter τ_{im} differs from the step difficulty parameter δ_{im} in that the step deviation is a deviation from the item location to the step difficulty ($\tau_{im} = \delta_{im} - \beta_i$). Since a sum of step deviations is constrained to be zero for each item ($\sum_{m=1}^{M_i} \tau_{im} = 0$) for model identification, $M_i - 1$ step deviation parameters are estimated for each item.

This twofold item parameterization is adopted in the ConQuest software (Adams, Wu, & Wilson, 2012; Wu, Adams, Wilson, & Haldane, 2007), which employs the Multidimensional Random Coefficients Multinomial Logit (MRCML; Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003) models. The twofold item parameterization is useful to figure out item locations, scale thresholds, and step difficulties in polytomous items as well as to input item properties to account for such item characteristics, and it is also convenient to constrain item parameters for model identification.

When using this parameterization for the PCM, the design matrix $\boldsymbol{X}_p$ contains item predictors for the item locations and item-by-step predictors for the step deviations, and the corresponding item parameter vector $\boldsymbol{\beta}$ contains both item location and step deviation parameters $(\beta_1, \beta_2, \tau_{11}, \tau_{21})$. The linear predictor vector of the PCM with the twofold item parameterization corresponding to Equation 3.6 can be written as

$$\begin{bmatrix} \eta_{p10} \\ \eta_{p10} + \eta_{p11} \\ \eta_{p10} + \eta_{p11} + \eta_{p12} \\ \eta_{p20} \\ \eta_{p20} + \eta_{p21} \\ \eta_{p20} + \eta_{p21} + \eta_{p22} \end{bmatrix} = \theta_p \times \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 2 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} \tau_{11} \\ \tau_{21} \end{bmatrix}. \tag{3.7}$$

In this illustration, in order to show different types of predictors and their effects, the design matrix $\boldsymbol{X}_p$ and the item parameter vector $\boldsymbol{\beta}$ are divided into two parts respectively: the item predictors and corresponding item location parameters which represent item main effects (Tuerlinckx & Wang, 2004), and the item-by-step predictors and corresponding step deviation parameters which represent item-by-step interaction effects. For each item with three response categories, elements of the item predictors for the item locations have values of category scores and one step deviation parameter is estimated due to the model constraint ($\tau_{i1} + \tau_{i2} = 0$). For example, for the first item,
when $m = 0$, $\eta_{p10} = 0$,
when $m = 1$, $\eta_{p10} + \eta_{p11} = 0 + (\theta_p - \beta_1 - \tau_{11}) = \theta_p - \beta_1 - \tau_{11}$, and
when $m = 2$, $\eta_{p10} + \eta_{p11} + \eta_{p12} = 0 + (\theta_p - \beta_1 - \tau_{11}) + (\theta_p - \beta_1 - \tau_{12}) = 2\theta_p - 2\beta_1$.

In addition, consider rating scale items in which item responses are ordinally scored to represent varying degrees of the attitude level. If we expect that the relative difficulties of the steps between categories, namely thresholds, should not vary across all items ($\tau_{im} = \tau_m$) and the number of steps are the same for all items ($M_i = M$), item location parameters β_i and common scale threshold parameters τ_m can be estimated (Masters & Wright, 1984):

$$\eta_{pim} = ln\frac{Pr(y_{pi} = m|\theta_p)}{Pr(y_{pi} = m-1|\theta_p)} = \theta_p - \beta_i - \tau_m\,, \qquad m = 1, \dots, M\,, \tag{3.8}$$

where $\theta_p \sim N(0, \sigma_\theta^2)$, $\eta_{pi0} = 0$, $\tau_0 = 0$, and $\sum_{m=1}^{M} \tau_m = 0$. This is the Rating Scale Model (RSM; Andrich, 1978) expressed as a special case of the PCM. The linear predictor vector for person *p* corresponding to Equation 3.8 is illustrated as follows:

$$\begin{bmatrix} \eta_{p10} \\ \eta_{p10} + \eta_{p11} \\ \eta_{p10} + \eta_{p11} + \eta_{p12} \\ \eta_{p20} \\ \eta_{p20} + \eta_{p21} \\ \eta_{p20} + \eta_{p21} + \eta_{p22} \end{bmatrix} = \theta_p \times \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 2 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \times [\tau_1]. \tag{3.9}$$

In the design matrix $\boldsymbol{X}_p$ and the item parameter vector $\boldsymbol{\beta}$ for the RSM, there are the item predictors and corresponding item location parameters which represent item main effects, and the step predictors and corresponding scale threshold parameters which represent step main effects (Tuerlinckx & Wang, 2004). When items have three response categories, one scale threshold parameter is estimated due to the model constraint ($\tau_1 + \tau_2 = 0$), and elements of the item predictors for the item locations have values of category scores. For example, for the first item,
when $m = 0$, $\eta_{p10} = 0$,
when $m = 1$, $\eta_{p10} + \eta_{p11} = 0 + (\theta_p - \beta_1 - \tau_1) = \theta_p - \beta_1 - \tau_1$, and
when $m = 2$, $\eta_{p10} + \eta_{p11} + \eta_{p12} = 0 + (\theta_p - \beta_1 - \tau_1) + (\theta_p - \beta_1 - \tau_2) = 2\theta_p - 2\beta_1$.

3.3. Item Explanatory Extensions of Polytomous Rasch Models under the Multivariate Generalized Linear Mixed Modeling Framework

3.3.1. Item Explanatory Extensions with Item Properties

To develop polytomous item explanatory IRT models, I will use the LLTM approach to extend the polytomous Rasch models by incorporating explanatory item properties as predictors in the design matrix $\boldsymbol{X}_p$ under the MGLMM framework. As discussed in the previous section, there are three types of predictors for the adjacent-categories logit based item response models, i.e. item predictors, step predictors, and item-by-step predictors. For item explanatory extensions, these predictors should be item properties for specific item characteristics. Corresponding to the types of predictors in the design matrix $\boldsymbol{X}_p$, the explanatory item properties such as item design variables can also be three specific types (Rijmen et al., 2003; Tuerlinckx & Wang, 2004): (a) item specific properties, unique factors inherent within items regardless of steps, (b) step specific properties, unique factors inherent within steps for all items, and (c) item-by-step specific properties, unique factors inherent within items and steps. These types of polytomous item properties are determined based on item design hypothesized by test developers.

Conceptually, it may be possible to consider the step specific properties. However, such properties are rare in practice. For example, ordered categorical levels in a learning progression framework are hypothesized by test developers (e.g. Table 3-1 in the first empirical study), and item steps between the hypothesized progression levels are hard to explain using a non-hypothetical external variable. In practice, it is tricky to take step specific properties into account in the adjacent-categories logit based item response models under the MGLMM framework. Due to the nature of adjacent-categories logits, the number of rows in the design matrix $\boldsymbol{X}_p$ is determined by the number of response categories for each item. And hence, the number of step specific parameters across all items would not likely be reduced by step specific properties, unless the number of response categories is changed. Moreover, it is difficult and controversial to interpret the effects of scale thresholds. Estimates of the scale thresholds indicate the common deviation of item steps from each item's location, which is difficult to relate to practical meaning about the relationship between item response categories, unless considered together with the item location. Because of these concerns, in this paper, I will leave out the case that step specific properties account for the scale thresholds in polytomous item explanatory models.

Accordingly, I will discuss how item specific properties and item-by-step specific properties are taken into account in the design matrix $\boldsymbol{X}_p$ to explain the item locations and the step difficulties respectively, by applying the LLTM approach to polytomous item response models. The LLTM (Fischer, 1973) is well known as an item explanatory IRT model for dichotomous item responses, in which the item difficulty parameters are decomposed into weighted sums of basic parameters, representing the effects of item properties on the item difficulties. I will use the LLTM approach as a general item explanatory extension and incorporate the two types of polytomous item properties into the PCM and the RSM in order to develop polytomous item explanatory models.

Fischer (1977) suggested this idea of a polytomous extension of the LLTM, and Fischer and Parzer (1991) examined an application of the LLTM approach to the RSM, and later Fischer

and Ponocny (1994) applied it to the PCM. These studies were based on conditional maximum likelihood (CML) estimation rather than marginal maximum likelihood (MML) estimation employed in the MGLMM framework. In particular, the item parameterization in the LLTM approach to the PCM (Fischer & Ponocny, 1994) looks complicated. They used a normalization constant c , basic parameters α_k, and given weights or values of item properties w_{imk}, so that $\delta_{im} = \sum_{k=1}^{K} \alpha_k w_{imk} + mc$. However, this item parameterization with the normalization constant requires relatively a large number of parameters to be estimated for each given weight, and makes them difficult to interpret. Those item parameters are not straightforward to calculate them, and even make it difficult to interpret the effects of item properties in terms of the ordinary PCM parameters such as the step difficulties. As Masters (1982) has pointed out, there is difficulty in Fischer's item parameterization including item properties for the polytomous extension of the LLTM.

In addition, Glas and Verhelst (1989) also investigated several extensions of the PCM using the MML estimation. Although they introduced a re-parameterization that $\eta_{im} = \sum_{h=1}^{m} \delta_{ih}$ to impose linear restrictions on the item parameters, they didn't address how to input item properties and interpret them in detail. Their approach necessarily requires a translation of the results into the original parametrization proposed by Masters (1982) but this translation is not straightforward, and hence it is hard to interpret item parameter estimates directly.

Meanwhile, Linacre (1989) proposed the Many-Facet Rasch Model (MFRM) with several maximum likelihood estimation methods. This model is actually an extension of the linear logistic model to polytomous item response data using the twofold item parameterization (Wu et al., 2007). Although the MFRM uses factorial notations with different subscripts for each facet[6] in model specification, it is essentially an additive linear logistic model based on a linear combination of the effects of multiple facets and a logistic transformation of ratios of successive category probabilities. In addition, the MFRM is a special case of the LLTM in that we can consider elementary components as facets of influential factors in item response process (Eckes, 2009; Rost, 2001; Rost & Langeheine, 1997). This approach is useful and flexible to form a measurement model which reflects various measurement situations such as examinee, items, raters, tasks, interviewers, or scoring criteria (Eckes, 2009; Linacre, 1989). For example, when raters observe a student's performance on tasks and then scores it to a response category in performance assessments, the students, raters, and tasks are facets which have an influence on the measurement outcomes. Despite the usefulness and flexibility of the MFRM approach, however, it has been mainly used to investigate the effects of facets such as rater/judge effects (e.g., Bonk & Ockey, 2003; Draney & Wilson, 2008; Eckes, 2009), but rarely used to examine the effects of item properties. In practice, item properties can be considered as sub-facets within the item facet (Wang & Wilson, 2005), and the MFRM approach can be used for item explanatory modeling.

[6] Widely defined, a facet can be any factor, variable, or component of the measurement situation that is assumed to influence test scores in a systematic way (Eckes, 2009; Linacre, 1989). Individual items and item properties, i.e., unique characteristics inherent in a set of items, can be considered as an item facet.

3.3.2. Many-Facet Rasch Model (MFRM) with a Decomposition of the Item Location Parameters

As discussed above, item parameterization incorporating item properties is complicated and difficult. To avoid such complication and difficulty, it is helpful to clarify the target difficulty parameters of polytomous items which are explained by item properties as well as the types of item properties which are incorporated in the design matrix $\boldsymbol{X}_p$. In the polytomous Rasch models based on the adjacent-categories logits under the MGLMM framework, there are two cases for polytomous item explanatory models: (a) the item location (overall item difficulty) parameters β_i can be explained by item specific properties, which represent item specific property effects, and (b) the step difficulty parameters δ_{im} can be explained by item-by-step specific properties, which represent item-by-step specific property effects. By default, I will consider a two-facet measurement situation (person and item) and the same set of item properties for each model, but this may not be the only consideration. Future studies could concern other-facet situations with different facet properties such as raters and criteria.

First, consider the former case in terms of the PCM. Note that the item location parameter is interpreted as the overall item difficulty for each polytomous item. As was done in the LLTM, one may want to account for the overall item difficulties by incorporating item specific properties into the design matrix $\boldsymbol{X}_p$. The twofold item parameterization (i.e. item location and step deviation) for the PCM enables us to impose linear restrictions focused on the item location parameters and also to estimate the step deviation parameters for each item. Thus, the restricted item location parameters β_i' will now be decomposed into weighted sums of item specific property effect parameters γ_k as follows:

$$\beta_i' = \sum_{k=0}^{K} \gamma_k x_{ik}\,, \qquad k = 0,\ \ldots,\ K \tag{3.10}$$

so that

$$\eta_{pim} = ln\frac{Pr\left(y_{pi} = m\middle|\theta_p\right)}{Pr\left(y_{pi} = m-1\middle|\theta_p\right)} = \theta_p - \sum_{k=0}^{K} \gamma_k x_{ik} - \tau_{im}, \tag{3.11}$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$, $\eta_{pi0} = 0$, $\tau_{i0} = 0$, and $\sum_{m=1}^{M_i} \tau_{im} = 0$ $(m = 1,\ \ldots,\ M_i)$. Here, γ_0 is the item intercept representing the difficulty for items with all x_{ik} = 0 for k > 0, γ_k is the regression weight or the effect of item property k on the overall item difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, x_{ik} is the value of item i on item property k, and τ_{im} is a step deviation parameter for the m-th step of item i. For model identification, the mean person ability is constrained to zero. Note that values of the step deviation parameter τ_{im} will be different from those in the PCM because the constructed item locations β_i' are calculated by weighted sums of the item specific property effects rather than directly estimated. In this model, the constructed step difficulties δ_{im}' can be calculated using $\delta_{im}' = \beta_i' + \tau_{im}$, where β_i' is the constructed item location for item i from Equation 3.10.

This polytomous item explanatory model is a variation of Linacre's (1989) MFRM. By using the twofold item parameterization, the item location parameters are decomposed into a linear

combination of the effects of sub-facets (i.e. item properties) in the item facet. The step deviation parameters are additively estimated for each item, because each item can have a different number of steps across ordered response categories in its rating scale structure. Building on the original idea of the MFRM approach, this item explanatory extension of the PCM will be called the "Many-Facet Rasch Model with a decomposition of the item location parameters".

Although this model is a variation of the polytomous MFRM, compared to factorial (ANOVA-style) notations which are commonly used in the MFRM, the model specification in Equation 3.11 has more general regression notations. While the original MFRM approach uses only categorical predictors to represent the effect of each facet (facet descriptive), this model can incorporate both categorical and continuous explanatory predictors inherent in the item facet to see the effect of item properties (item explanatory). For instance, this model is useful and suitable when one may want to see the effects of passage lengths and contents on the overall item difficulties in a reading comprehension test. Moreover, the step deviation parameters can be estimated variously according to the scale structure in the MFRM, but they are estimated for each item in this model.

To show a simple illustration of this MFRM variation under the MGLMM, suppose we have two polytomous items with three response categories for each ($M_1 = M_2 = 2$), and one item property of item format (X and Y types). The item property is categorical in this example, but it can be continuous. Also, suppose that two items are identified as two types of the item format in the observed data: X for Item 1 and Y for Item 2. After using dummy coding[7] for the categorical item property (reference for Y), an intercept (γ_0) and the effect of item format X (γ_X) are estimated to account for the overall item difficulties. In addition, two step deviation parameters (τ_{11}, τ_{21}) are estimated for the two items. Again, note that the linear predictor vector $\boldsymbol{\eta}_p$ contains a cumulative sum of the linear predictor elements η_{pim} of the adjacent-categories logits for all category scores in each item.

First, consider the linear predictor vector for Item 1 which has the X type of item format:
when $m = 0$, $\eta_{p10} = 0$,
when $m = 1$, $\eta_{p10} + \eta_{p11} = 0 + (\theta_p - \gamma_0 - \gamma_X - \tau_{11}) = \theta_p - \gamma_0 - \gamma_X - \tau_{11}$, and
when $m = 2$, $\eta_{p10} + \eta_{p11} + \eta_{p12} = 0 + (\theta_p - \gamma_0 - \gamma_X - \tau_{11}) + (\theta_p - \gamma_0 - \gamma_X - \tau_{12}) = 2\theta_p - 2\gamma_0 - 2\gamma_X$, due to $\tau_{11} + \tau_{12} = 0$.

Second, consider the linear predictor vector for Item 2 which has the Y type of item format:
when $m = 0$, $\eta_{p20} = 0$,
when $m = 1$, $\eta_{p20} + \eta_{p21} = 0 + (\theta_p - \gamma_0 - \tau_{21}) = \theta_p - \gamma_0 - \tau_{21}$, and
when $m = 2$, $\eta_{p20} + \eta_{p21} + \eta_{p22} = 0 + (\theta_p - \gamma_0 - \tau_{21}) + (\theta_p - \gamma_0 - \tau_{22}) = 2\theta_p - 2\gamma_0$, due to $\tau_{21} + \tau_{22} = 0$.

Consequently, the linear predictor vector for person p of this MFRM variation corresponding to Equation 3.11 can be written as

[7] Although it is possible to use effect coding for categorical item properties, dummy coding is used for common interpretation of the item specific property effect parameters γ_k. Effect coding might be useful if an interaction between persons and items were considered in order to model fixed effects or random weights of the person-by-item predictors (see De Boeck & Wilson, 2004).

$$\begin{bmatrix} \eta_{p10} \\ \eta_{p10}+\eta_{p11} \\ \eta_{p10}+\eta_{p11}+\eta_{p12} \\ \eta_{p20} \\ \eta_{p20}+\eta_{p21} \\ \eta_{p20}+\eta_{p21}+\eta_{p22} \end{bmatrix} = \theta_p \times \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 0 & 0 \\ 1 & 0 \\ 2 & 0 \end{bmatrix} \times \begin{bmatrix} \gamma_0 \\ \gamma_X \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} \tau_{11} \\ \tau_{21} \end{bmatrix}. \quad (3.12)$$

In this linear predictor vector, the design matrix $\boldsymbol{X}_p$ and the item parameter vector $\boldsymbol{\beta}$ are divided into two parts respectively: the item specific properties (item predictors) and corresponding item specific property effect parameters (γ_0, γ_X), and the item-by-step predictors and corresponding step deviation parameters (τ_{11}, τ_{21}). In addition, we can calculate the constructed step difficulties for each item based on the estimated item parameters. For example, $\delta'_{11} = \beta'_1 + \tau_{11} = \gamma_0 + \gamma_X + \tau_{11}$, and $\delta'_{22} = \beta'_2 + \tau_{22} = \gamma_0 + (-\tau_{21})$.

3.3.3. Linear Partial Credit Model (LPCM) with a Decomposition of the Step Difficulty Parameters

The second case is an item explanatory extension of the PCM, in which the step difficulties are determined by item-by-step specific property effects. Compared to the effects of item specific properties in the former approach, the effects of item-by-step specific properties are step specific so that their estimates can vary across steps in this approach. For model specification, we can impose linear restrictions on the step difficulty parameters by incorporating item-by-step specific properties into the design matrix $\boldsymbol{X}_p$. Thus, the restricted step difficulty parameters δ'_{im} will now be decomposed into weighted sums of item-by-step specific property effect parameters ω_{km} as follows:

$$\delta'_{im} = \sum_{k=0}^{K} \omega_{km} x_{ik}\,, \qquad k = 0,\ \ldots,\ K \quad (3.13)$$

so that

$$\eta_{pim} = ln\frac{Pr(y_{pi} = m|\theta_p)}{Pr(y_{pi} = m-1|\theta_p)} = \theta_p - \sum_{k=0}^{K} \omega_{km} x_{ik}\,, \quad (3.14)$$

where $\theta_p \sim N(0, \sigma_\theta^2)$, $\eta_{pi0} = 0$, $\omega_{k0} = 0$, and $m = 1,\ \ldots,\ M_i$. Here, ω_{0m} is the step intercept representing the m-th step difficulty for items with all x_{ik} = 0 for k > 0, ω_{km} is the regression weight or the effect of item property k on the m-th step difficulties (i.e. the item property effect is step specific), x_{io} is the constant item predictor in which a value of 1 for all items, and x_{ik} is the value of item i on item property k. Note that the item property values of x_{ik} in the linear predictor element η_{pim} are expanded over all category scores for each item so that they can be incorporated as item-by-step specific properties in the design matrix $\boldsymbol{X}_p$ (see below a simple illustration for the linear predictor vector). For model identification, the mean person ability is constrained to zero. According to a configuration of item properties for each item, the constructed step difficulties δ'_{im} can be calculated by weighted sums of the estimated item property effects from Equation 3.13.

This polytomous item explanatory model is a variation of Fischer and Ponocny's (1994) Linear Partial Credit Model (LPCM). Although they used less interpretable item parameters such as a normalization constant and basic parameter, they imposed linear restrictions on the step difficulty parameters and their basic parameters can be regarded as the effects of item-by-step specific properties given weights or values of item properties. Building on the original idea of the LPCM, this item explanatory extension of the PCM will be called the "Linear Partial Credit Model with a decomposition of the step difficulty parameters".

Compared to the original parameterization for the LPCM, however, the model specification in Equation 3.14 employs the item-by-step specific property effect parameters ω_{km}, which are more interpretable item parameters for the effects of item properties. This model allows estimation of the item property effect parameters step-specifically for each step of item properties so we can interpret them as the effects of item properties on the step difficulties for each step. Tuerlinckx and Wang (2004) introduced this step-specific explanatory approach to the PCM and they also added person predictors. For the item explanatory part of the model, they specified item regression components for each step using the same set of item properties. Building on their work, I combined the item regression components using one formula in Equation 3.13 under the MGLMM framework. Moreover, while the original LPCM parameterization makes it intractable to incorporate item properties into the design matrix $\boldsymbol{X}_p$ for the adjacent-categories logit based item response models, this model makes it tractable.

On the other hand, this model can also be considered as a special or restricted version of the MFRM variation. By applying the twofold item parameterization, the effects of item property k on the step difficulties can be decomposed into the effects on the item locations and on the step deviations separately so that $\omega_{km} = \gamma_k + \tau_{km}$. In this case, this LPCM variation would be mathematically or technically similar to the MFRM approach except for additional restrictions on the step deviation parameters. However, the item property effects on the step deviations are not quite substantive and it is even hard to interpret them. Rather, the target difficulty parameters of polytomous items which are meaningfully explained by the two types of item properties are crucial to distinguish between the MFRM and the LPCM approaches.

For a simple illustration of this LPCM variation under the MGLMM, I will use the previous example for the MFRM variation, but now will use decomposition of step difficulty parameters instead of decomposition of item location parameters is considered. After using dummy coding for the categorical item property (reference for Y), an intercept (ω_{01},) and the effect of item format X (ω_{X1}) are estimated to account for the step difficulties of the first step, and the same set of other parameters (ω_{02}, ω_{X2}) are estimated to explain the step difficulties of the second step.

First, consider the linear predictor vector for Item 1 which has the X type of item format:
when $m = 0$, $\eta_{p10} = 0$,
when $m = 1$, $\eta_{p10} + \eta_{p11} = 0 + (\theta_p - \omega_{01} - \omega_{X1}) = \theta_p - \omega_{01} - \omega_{X1}$, and
when $m = 2$, $\eta_{p10} + \eta_{p11} + \eta_{p12} = 0 + (\theta_p - \omega_{01} - \omega_{X1}) + (\theta_p - \omega_{02} - \omega_{X2}) = 2\theta_p - \omega_{01} - \omega_{02} - \omega_{X1} - \omega_{X2}$.

Second, consider the linear predictor vector for Item 2 which has the Y type of item format:
when $m = 0$, $\eta_{p20} = 0$,
when $m = 1$, $\eta_{p20} + \eta_{p21} = 0 + (\theta_p - \omega_{01}) = \theta_p - \omega_{01}$, and
when $m = 2$, $\eta_{p20} + \eta_{p21} + \eta_{p22} = 0 + (\theta_p - \omega_{01}) + (\theta_p - \omega_{02}) = 2\theta_p - \omega_{01} - \omega_{02}$.

Accordingly, the linear predictor vector for person *p* of this LPCM variation corresponding to Equation 3.14 can be expressed as

$$\begin{bmatrix} \eta_{p10} \\ \eta_{p10} + \eta_{p11} \\ \eta_{p10} + \eta_{p11} + \eta_{p12} \\ \eta_{p20} \\ \eta_{p20} + \eta_{p21} \\ \eta_{p20} + \eta_{p21} + \eta_{p22} \end{bmatrix} = \theta_p \times \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} \omega_{01} \\ \omega_{02} \\ \omega_{X1} \\ \omega_{X2} \end{bmatrix}. \quad (3.15)$$

In this linear predictor vector, the design matrix $\boldsymbol{X}_p$ contains the item-by-step specific properties (item-by-step predictors) and the corresponding item parameter vector $\boldsymbol{\beta}$ contains the item-by-step specific property effect parameters (ω_{01}, ω_{02}, ω_{X1}, ω_{X2}). In addition, we can calculate the constructed step difficulties for each item based on the estimated item parameters. For example, $\delta'_{11} = \omega_{01} + \omega_{X1}$, and $\delta'_{22} = \omega_{02}$.

So far, I have discussed two item explanatory extensions of the PCM by using the LLTM approach under the MGLMM framework. The former is the MFRM with a decomposition of the item location parameters, and the latter is the LPCM with a decomposition of the step difficulty parameters. The two polytomous item explanatory models are conceptually and functionally different models. In particular, the target difficulty parameters of polytomous items which are explained, the types of item properties incorporated in the design matrix, and the types of item property effects are different between the two models. Compared to the effects of item specific properties on the overall item difficulties in the former model, the effects of item-by-step specific properties on the step difficulties in the latter model are step specific. For model selection between the two models, it is necessary to clarify which polytomous item difficulties are explained by item properties as well as to examine what kind of item properties are included in the design matrix.

When we impose linear restrictions on the item parameters, we can take not only the main effects of item properties into account but also the interaction effects between them. As for a regular regression model, interaction terms between the item properties can be added into the item explanatory components. In terms of the goodness-of-fit, polytomous item explanatory models with item property interactions may fit the data better because of richer predictors.

Furthermore, the MFRM approach can be simply applied to the RSM, resulting in the Linear Rating Scale Model (LRSM; Fischer & Parzer, 1991; Fischer, 1997). In a similar way to the MFRM variation, the item location parameters can be decomposed into weighted sums of item specific property effects and the scale threshold parameters are estimated as usual. For more details about the LRSM under the MGLMM framework, please see the Appendix A.1 of this paper. Although the LRSM can be specified, it is still difficult and controversial to interpret the scale threshold effects in the item explanatory measurement. In addition, the MFRM approach may have another interesting variation in which item specific properties and/or item-by-step specific properties account for the overall item difficulties and the step difficulties simultaneously. By employing the twofold item parameterization, the MFRM may be able to differentiate the item specific property effects from the item-by-step specific property effects on the step difficulties. A future study can consider this variation.

To sum up, the polytomous item explanatory models I have specified in this paper are special cases or restricted models of the PCM under the MGLMM framework. The step difficulty

parameters or the item location parameters are constrained to be a linear combination of item property effect parameters. This linear constraint enables us to estimate a smaller number of item parameters, compared to the PCM. This is a methodological advantage in extracting essential and meaningful elementary components from the item response data and thus it has explanatory value (De Boeck & Wilson, 2004; Kubinger, 2009).

3.4. Empirical Studies

3.4.1. Empirical Data Analysis

My research aim for the empirical studies is to see how the proposed polytomous item explanatory models under the MGLMM framework, the MFRM variation and the LPCM variation, work for practical measurement/assessment data sets. The two item explanatory approaches are conceptually and functionally different in terms of the target difficulty parameters of polytomous items to be explained and the types of incorporated item properties as well as the types of item property effects. To see practical difference between the two approaches, I fit both of the two models to the Carbon Cycle assessment data and to the verbal aggression data respectively. In addition, since the PCM is a saturated model for the polytomous Rasch models, it was fitted to each empirical data set to see how the two polytomous item explanatory models perform compared to the saturated model.

In order to fit the adjacent-categories logit based item response models under the MGLMM framework, I used the *gllamm* command in Stata (Rabe-Hesketh, Skrondal, & Pickles 2004). Since the MGLMM is a special case of the GLLAMM (Zheng & Rabe-Hesketh, 2007), all models under the MGLMM framework can be fit by the *gllamm* command which is implemented with marginal maximum likelihood (MML) estimation. For model identification, the mean person ability was constrained to be zero for all the fitted models in the empirical studies.

Regarding the goodness-of-fit of the models, the Likelihood Ratio test (LR test; Andersen, 1973) was conducted to compare the nested models. The two polytomous item explanatory models are nested within the PCM, and the LPCM variation is nested within the MFRM variation due to more restrictions on the item parameters. Three goodness-of-fit indices were also reported for each model: the deviance (D), the Akaike Information Criterion (AIC; Akaike, 1974), and the Bayesian Information Criterion (BIC; Schwarz, 1978). Based on the deviance, the AIC and BIC were computed with a penalty included for the number of estimated parameters and observations (De Boeck & Wilson, 2004):

$$D = -2 \times Ln(Likelihood),$$
$$\text{AIC} = D + 2k,$$
$$\text{BIC} = D + Ln(N) \times k,$$

where k is the number of estimated parameters and N is the number of observations. These indices can be viewed as measures composed of model fit (accuracy) and complexity (parsimony) of the fitted model (Wagenmakers & Farrell, 2004). Given two models fit on the same data, the smaller value of these indices represents the model is better fitted to the data statistically.

In addition to the goodness-of-fit comparison, a graphical comparison and a correlation analysis were conducted to see if there is agreement for the estimated and calculated step difficulties between the fitted models (see Fischer, 1973; Poinstingl, 2009). In particular, the estimated step difficulties in the PCM were compared with the constructed step difficulties which were calculated by weighted sums of the estimated item property effects from the two polytomous item explanatory models.

Lastly, in each empirical study, the item property effects found from the MFRM variation and the LPCM variation were examined to see how the observed item properties can explain and predict the difficulties of polytomous items—the overall item difficulties or the step difficulties—differently through the two models.

3.4.2. Application to the Carbon Cycle Assessment data

Data

A subset of the Carbon Cycle assessment data, specifically pre-test responses of the Math and Science Partnership (MSP) Carbon student assessment data collected during the 2010-2011 academic year, was used in the first empirical study. Participants were middle and high school students from urban, suburban, and rural area in five states (CA, CO, MD, MI, NY).

Table 3-1. The Learning Progression Framework for the Carbon Cycle assessment

Level	General description	Exemplary response
4	Linking processes with matter and energy as constraints	The mass decrease because the mold uses energy to grow it goes through cellular respiration, the result of cellular respiration is CO2 and H20
3	Changes of molecules and energy forms with unsuccessful constraints	The mold decays the bread and turns it into energy so the mass decreases
2	Force-dynamic accounts with hidden mechanisms	Even though the mold is converting the bread into biomass, the mold also has gained mass through photosynthesis.
1	Macroscopic force-dynamic accounts	The mass increases because the mold adds its weight to the original mass of the bread

In the Carbon Cycle project (Mohan, Chen, & Anderson, 2009), the assessment and items were developed based on a learning progression framework for carbon cycling in socio-ecological systems in science education (e.g. Jin et al., 2015), by conducting an iterative process of designing, analyzing, and modifying assessments with interviews for secondary school students. The learning progression for carbon cycling in socio-ecological systems has four ordered levels of achievement corresponding to students' progress toward more sophisticated reasoning about biogeochemical

processes (Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009), as shown in Table 3-1. On the basis of the Carbon Cycle learning progression framework, items were developed to ask students to answer forced choice questions and explain their choices, and their scoring rubrics were developed to score item responses into the four ordered achievement levels.

The 13 Carbon Cycle items consist of a combination of three categorical item properties: (a) the Process property has three types of biogeochemical processes in carbon cycling that transform carbon in socio-ecological systems at multiple scales—cellular respiration (CR), photosynthesis (PS), and digestion/biosynthesis (DB), (b) the Progress property has four learning progress variables related to the carbon cycling process—large-scale systems (LS), micro-scale systems (MS), energy (EN), and mass (MA), and (c) the Format property has two item response formats—multiple choice with explanation (MC) and yes/no choice with explanation (YN). For example, the first item (BODYTEMP) can be identified as a combination of a CR predictor in the Process property, a EN predictor in the Progress property, and a MC predictor in the Format property. The text of the first item is shown below:

Q1. [BODYTEMP] Your body produces heat to maintain its normal temperature. Where does the heat mainly come from? Please choose the ONE answer that you think is best and explain why you think that the answer you chose is better than the others. If you think some of the other answers are also partially right, please explain why you think so.

A. The heat mainly comes from sunlight.
B. The heat mainly comes from the clothes you are wearing.
C. The heat mainly comes from the foods you eat.
D. The heat mainly comes from your body when you are exercising.

Table 3-2. The Carbon Cycle items composed of three item properties

Item		Process			Progress				Format	
Num.	Name	Cellular Respiration (CR)	Photosyn-thesis (PS)	Digestion/ Biosynthesis (DB)	Large Scale (LS)	Micro Scale (MS)	Energy (EN)	Mass (MA)	Multiple Choice (MC)	Yes/No Choice (YN)
1	BODYTEMP	1	0	0	0	0	1	0	1	0
2	CARBCYC2	1	0	0	1	0	0	0	0	1
3	CARBPLNT	0	1	0	0	1	0	0	0	1
4	CONTCARB	0	0	1	0	1	0	0	0	1
5	CRKTGRWTH	1	0	0	0	0	0	1	1	0
6	EATAPPLE	0	0	1	0	1	0	0	0	1
7	INFANT	0	0	1	0	0	0	1	0	1
8	LIGHTEN	0	1	0	0	0	1	0	1	0
9	LIGHTEN2	0	1	0	0	0	1	0	1	0
10	MAPLEMASS	0	1	0	0	0	0	1	1	0
11	PLNTGSENS	1	0	0	0	1	0	0	0	1
12	PLNTGRWTH	0	1	0	0	0	0	1	0	1
13	THINGTREE	0	1	0	0	0	0	1	0	1

Through item analyses for the data and consideration of item design, the items can be classified by the predictors of each item property as in Table 3-2. These predictors of the item properties are functioning as weights of the elementary components which are gathered into a Q matrix in the LLTM approach (Kubinger, 2009; Poinstingl, 2009).

The items were polytomously scored by the scoring rubrics developed based on the learning progression framework. Four levels of student achievement were assigned to score it corresponding to the learning progression in the Carbon Cycle assessment. Figure 3-1 shows the scoring rubric of the first item. For instance, students in the highest level are hypothesized to be able to correctly link the carbon-transforming processes of cellular respiration with energy and matter as constraints. Students' responses reaching this scoring criterion are scored at Level 4. In a similar manner, this scoring rubric works for the other levels. If students' answers are not related to the question, or are illegible or nonsensical, they are treated as missing.

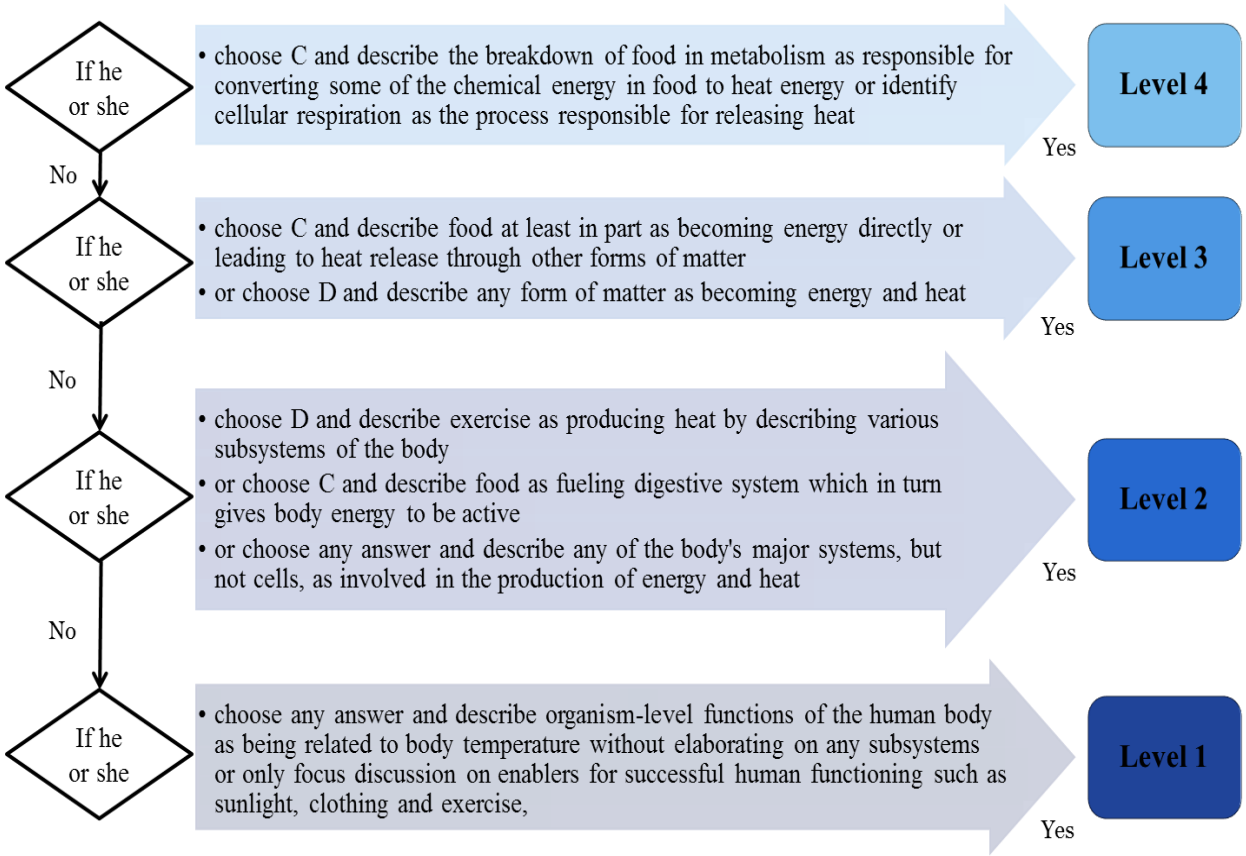


Figure 3-1. Flow chart of the scoring rubric for the item 1 (BODYTEMP)

In the initial data, some items (e.g. CARBCYC2, CONTCARB, CRKTGRWTH, MAPLEMASS, PLNTGSENS, THINGTREE) had too few or zero responses for the highest level, which would result in poor or unfeasible estimation of the third step difficulty parameters. In this empirical example, item responses of the four level categories (1, 2, 3, 4) were recoded into three category scores (0, 1, 2, 2) to avoid this sparse responses problem. We can interpret the second step difficulties as the relative difficulties as one goes from the Level 2 to the combined Levels of 3 and 4. Also, some cases which have less than three valid item responses were deleted to enhance

the quality of estimation and reduce the error of estimation for person abilities. After refining the data and deleting missing values, in total, 21,417 observations[8] of 1,157 students on the 13 polytomous Carbon Cycle items were included in the final data set.

In addition, the three categorical item properties were dummy coded so that the digestion/ biosynthesis (DB) in the Process property, the mass (MA) in the Progress property, and the yes/no choice with explanation (YN) in the Format property were a reference for each item property. The item properties would act as item specific properties or item-by-step specific properties based on their hypothesized effects on the item locations or the step difficulties. If they are hypothesized to be different across the items but constant across their steps, they would be regarded as item specific properties. If they are hypothesized to be differentiated across the items and their steps, they would be considered as item-by-step specific properties rather than item specific properties. In this empirical study, both types of item properties were taken into account for applications of the two proposed polytomous item explanatory models to the Carbon Cycle assessment data.

Results

According to the data analysis procedure, the PCM, the MFRM variation (also called the MFRM here), and the LPCM variation (also called the LPCM here) were fit to the Carbon Cycle assessment data. Table 3-3 shows the results of the fitted models to see performance of the two polytomous item explanatory models that I proposed compared to the saturated model, the PCM.

The estimated person variance (σ_θ^2) was 0.81 for the MFRM and 0.78 for the LPCM, which were smaller than 1.17 in the PCM. This indicates shrinkage of the estimated person variance in the polytomous item explanatory models as in the LLTM. This can be understood as a scaling effect that comes from the deficient explanation of the item parameters by imposing linear restrictions of the item properties (De Boeck & Wilson, 2004). In terms of the person variance shrinkage, the LPCM was slightly more shrunk than the MFRM but the effects were very similar between the two models.

In terms of the goodness-of-fit, both the MFRM and the LPCM appeared to fit worse than the PCM. The LR test comparing to the PCM was significant for both the MFRM ($\chi^2(6) = 767.12$, $p < .001$) and the LPCM ($\chi^2(12) = 971.97$, $p < .001$). This result was confirmed in that all goodness-of-fit indices of the two models were greater than the PCM. This is as expected because the item explanatory models commonly fit worse than the saturated model (Kubinger, 2009; Tuerlinckx & Wang, 2004). A methodological advantage of a smaller number of item parameters in the LLTM approach is at the cost of statistically lower goodness-of-fit.

When comparing between the two item explanatory models, both the AIC and BIC indicate that the MFRM showed a superior goodness-of-fit than the LPCM to the Carbon Cycle data. The LR test comparing the LPCM to the MFRM was also significant as $\chi^2(6) = 204.85$ ($p < .001$). This result makes sense because the step deviation parameters were freely estimated for each item in the MFRM and hence it had a larger number of the estimated parameters than the LPCM. Nonetheless, this statistical result doesn't mean that the LPCM is methodologically or practically inferior to the MFRM. They are conceptually and functionally different models in that the target

[8] When using the *gllamm* command, item response data are long-formatted and expanded over all category scores for each item. Since all items were refined to have three response categories in the final data set, there were 21,417 (7,139 × 3) observations, after deleting missing values.

difficulty parameters of polytomous items which are explained, the types of incorporated item properties, and the types of item property effects are different.

Table 3-3. Performance of the MFRM and the LPCM on the Carbon Cycle assessment data

		Partial Credit Model (PCM)	Many-Facet Rasch Model (MFRM)	Linear Partial Credit Model (LPCM)
Step difficulty				
Item 1	δ_{11}	0.23	-0.02	-0.14
	δ_{12}	2.16	1.53	1.46
Item 2	δ_{21}	-4.02	-4.06	-4.05
	δ_{22}	2.90	2.59	2.59
Item 3	δ_{31}	-1.78	-1.85	-1.33
	δ_{32}	1.55	1.31	0.98
Item 4	δ_{41}	-1.64	-1.65	-1.67
	δ_{42}	1.79	1.55	1.58
Item 5	δ_{51}	-0.37	-0.29	0.22
	δ_{52}	2.39	3.21	2.30
Item 6	δ_{61}	-2.61	-1.51	-1.67
	δ_{62}	0.93	1.41	1.58
Item 7	δ_{71}	-0.59	-0.71	-0.55
	δ_{72}	2.81	1.88	1.59
Item 8	δ_{81}	-0.76	-0.76	-0.47
	δ_{82}	0.77	0.61	0.42
Item 9	δ_{91}	-0.66	-0.39	-0.47
	δ_{92}	0.20	0.24	0.42
Item 10	δ_{101}	0.94	0.56	-0.11
	δ_{102}	1.57	0.70	1.26
Item 11	δ_{111}	-0.44	-0.42	-1.01
	δ_{112}	1.87	1.54	2.02
Item 12	δ_{121}	0.81	0.15	-0.21
	δ_{122}	2.33	0.58	1.00
Item 13	δ_{131}	-2.51	-0.44	-0.21
	δ_{132}	0.03	1.18	1.00
Variance				
	σ_θ^2	1.17	0.81	0.78
Goodness-of-fit				
	D	12223.62	12990.74	13195.59
	AIC	12277.62	13032.74	13225.59
	BIC	12492.86	13200.15	13345.17
	k	27	21	15

Note: k = The number of estimated parameters.

In addition to the goodness-of-fit comparison, a graphical comparison is useful to see performance of the two polytomous item explanatory models, by examining agreement for the estimated and calculated step difficulties between the fitted models. The step difficulties δ_{im} for each item were directly estimated in the PCM, and the constructed step difficulties δ'_{im} were calculated from the estimated item property effects in each item explanatory model. The PCM was used as the reference model for graphical model comparisons. Figure 3-2 shows that the MFRM had similar agreement to the LPCM although a few more step difficulty points were close to the 45-degree line which represents perfect alignment. However, the agreement gap in the two models was quite notable. In both models, the first step difficulty of item 13 was located farthest from the 45-degree line.

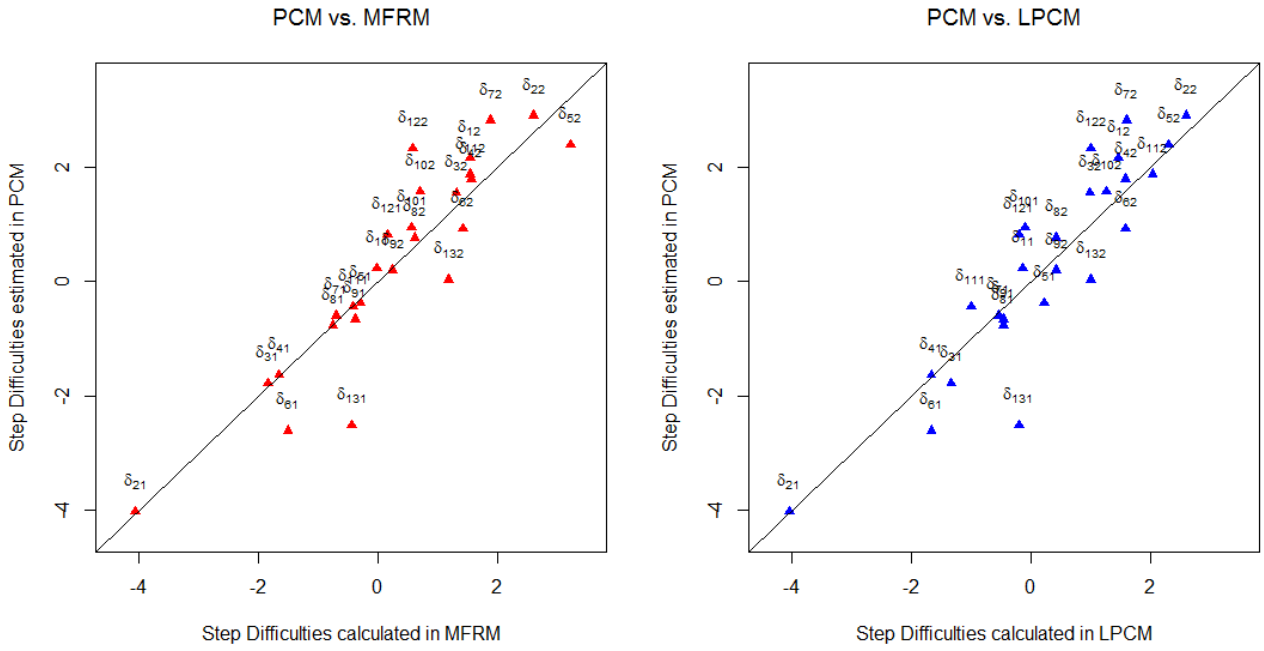


Figure 3-2. Graphical comparison of step difficulties between the models fitted to the Carbon Cycle assessment data

A correlation analysis in Table 3-4 shows the same results to the graphical comparison. Correlations with the PCM were similar in the two item explanatory models (ρ = 0.913 for the MFRM, and ρ = 0.905 for the LPCM). A correlation of the calculated step difficulties between the two item explanatory models (ρ = 0.970) was higher than their correlations with the estimated step difficulties in the PCM, but it was not complete agreement. This implies that the MFRM and the LPCM could predict the step difficulties in a similar but different manner. Despite the gap of goodness-of-fit between the two item explanatory models, these results showed comparable performance in reconstructing the step difficulties using the item property effects.

Table 3-4. Correlations of the estimated and calculated step difficulties in the Carbon Cycle items

	Partial Credit Model (PCM)	Many-Facet Rasch Model (MFRM)	Linear Partial Credit Model (LPCM)
PCM	1	0.913	0.905
MFRM		1	0.970

Table 3-5. Item property effects on the Carbon Cycle items in the MFRM variation

Predictor	Item Parameter	Estimate	Standard Error	P-value
Intercept	γ_0	0.59	0.07	< .001
Process (cellular respiration)	γ_{CR}	0.61	0.07	< .001
Process (photosynthesis)	γ_{PS}	-0.22	0.06	< .001
Progress (large-scale systems)	γ_{LS}	-1.93	0.21	< .001
Progress (micro-scale systems)	γ_{MS}	-0.64	0.06	< .001
Progress (energy)	γ_{EN}	-0.71	0.07	< .001
Format (multiple choice)	γ_{MC}	0.26	0.08	< .001
Step deviation parameters for each item	τ_{11}	-0.77	0.10	< .001
	τ_{21}	-3.33	0.20	< .001
	τ_{31}	-1.58	0.07	< .001
	τ_{41}	-1.60	0.10	< .001
	τ_{51}	-1.75	0.11	< .001
	τ_{61}	-1.46	0.10	< .001
	τ_{71}	-1.29	0.10	< .001
	τ_{81}	-0.68	0.10	< .001
	τ_{91}	-0.32	0.09	0.001
	τ_{101}	-0.07	0.10	0.474
	τ_{111}	-0.98	0.10	< .001
	τ_{121}	-0.21	0.09	0.023
	τ_{131}	-0.81	0.09	< .001

In order to focus on practical differences between the two polytomous item explanatory models, the effects of the three item properties on the overall item difficulties or the step difficulties were reported for each model. Table 3-5 shows the results of the item property effects in the

MFRM variation. Recall that there are three categorical item properties (Process, Progress, and Format) in the design of the Carbon Cycle assessment. The three item properties incorporated as item specific properties were taken into account to explain and predict the overall item difficulties (item locations). As we examined the item specific property effect parameters γ, the contributions of all the item properties were statistically significant (but the step deviation parameters τ are not of interest for interpretation).

For the Process property, holding other properties constant, cellular respiration (CR; γ_{CR} = 0.61, $p < .001$) made the overall item difficulty of an item more difficult and photosynthesis (PS; γ_{PS} = -0.22, $p < .001$) made them less difficult than digestion/biosynthesis (reference). In the Progress property, keeping other properties constant, compared to mass (reference), large-scale systems (LS; γ_{LS} = -1.93, $p < .001$), energy (LS; γ_{EN} = -0.71, $p < .001$), and micro-scale systems (MS; γ_{MS} = -0.64, $p < .001$) made the overall item difficulty of an item less difficult in order. For the Format property, holding other properties constant, multiple choice with explanation formatted items (MC; γ_{MC} = 0.26, $p < .001$) had greater overall item difficulties than yes/no choice with explanation formatted items (reference).

In addition, 13 step deviation parameters were estimated for each item. Accordingly, the constructed step difficulties for each observed combination item could be calculated by using the estimated item property effects and step deviation parameters. For example, the second step difficulty of the item 1 (BODYTEMP), δ'_{12} was calculated as 0.59 (γ_0) + 0.61 (γ_{CR}) − 0.71 (γ_{EN}) − 0.26 (γ_{MC}) + 0.77 ($-\tau_{11}$) ≈ 1.53.

Table 3-6. Item property effects on the Carbon Cycle items in the LPCM variation

Predictor	Item Parameter	Estimate	Standard Error	P-value
Intercept	ω_{01}	-0.55	0.09	< .001
	ω_{02}	1.59	0.10	< .001
Process (cellular respiration)	ω_{CR1}	0.66	0.11	< .001
	ω_{CR2}	0.45	0.12	< .001
Process (photosynthesis)	ω_{PS1}	0.34	0.09	< .001
	ω_{PS2}	-0.60	0.09	< .001
Progress (large-scale systems)	ω_{LS1}	-4.16	0.36	< .001
	ω_{LS2}	0.55	0.23	0.016
Progress (micro-scale systems)	ω_{MS1}	-1.12	0.08	< .001
	ω_{MS2}	-0.01	0.09	0.885
Progress (energy)	ω_{EN1}	-0.36	0.09	< .001
	ω_{EN2}	-0.84	0.13	< .001
Format (multiple choice)	ω_{MC1}	0.10	0.10	0.310
	ω_{MC2}	0.26	0.14	0.057

Table 3-6 shows the results of the item property effects in the LPCM variation. The three item properties incorporated as item-by-step specific properties were taken into account so that they could explain and predict the step difficulties for each step. Based on the estimated item-by-step specific property effect parameters ω, we can see how the three item properties affect the step difficulties step-specifically in the Carbon Cycle assessment. Recall that the first step's difficulties are interpreted as the relative difficulties as one goes from the Level 1 to the Level 2, and the second step's difficulties represent the relative difficulties as one goes from the Level 2 to the combined Levels of 3 and 4.

With regard to the effects of the three item properties on the step difficulties of the first step, keeping other properties constant, cellular respiration process (CR; ω_{CR1} = 0.66, p < .001) in the Process property, mass progress variable (reference) in the Progress property, and multiple choice with explanation format (MC; ω_{MC1} = 0.10, p = 0.310) in the Format property made the items more difficult than others within each item property, as a student answered from the Level 1 to the Level 2. In contrast, digestion/biosynthesis process (reference) in the Process property, large-scale systems progress variable (LS; ω_{LS1} = -4.16, p < .001) in the Progress property, and yes/no choice with explanation format (reference) in the Format property made the step difficulties less difficult than others within each item property.

For the step difficulties of the second step, holding other properties constant, cellular respiration process (CR; ω_{CR2} = 0.45, p < .001) in the Process property, large-scale systems progress variable (LS; ω_{LS2} = 0.55, p = 0.016) in the Progress property, and multiple choice with explanation format (MC; ω_{MC2} = 0.26, p = 0.057) in the Format property made it more difficult than others within each item property for a student to achieve either of the Level 3 and Level 4 from the Level 2 in the learning progression for carbon cycling. In contrast, photosynthesis process (PS; ω_{PS2} = -0.60, p < .001) in the Process property, energy progress variable (EN; ω_{EN2} = -0.84, p < .001) in the Progress property, and yes/no choice with explanation format (reference) in the Format property made it less difficult than others within each item property.

Furthermore, we could reconstruct the step difficulties for each observed combination item by weighted sums of the estimated item property effects. For example, the second step difficulty of the item 1 (BODYTEMP), δ'_{12} was calculated as $1.59\ (\omega_{02}) + 0.45\ (\omega_{CR2}) - 0.84\ (\omega_{EN2}) + 0.26\ (\omega_{MC2})$ = 1.46. We could also predict the step difficulties for the unobserved combination items in the Carbon Cycle data. For instance, the first step difficulty for the (CR, MA, YN) combination item which was not observed in the current data set could be predicted as $\delta'_{(CR,MA,YN)1} = -0.55\ (\omega_{01}) + 0.66\ (\omega_{CR1})$ = 0.11.

In sum, the MFRM and the LPCM worked effectively for the Carbon Cycle data in a similar but different manner. The two models showed the comparable performance in reconstructing the step difficulties using the estimated item property effects. From a superior goodness-of-fit of the MFRM to the LPCM, one could say the three item properties explained the overall item difficulties better than the step difficulties. However, this might not be the case. These two polytomous item explanatory models are methodologically and practically different models. In practice, we could figure out how the three item properties (Process, Progress, and Format) affect the overall item difficulties in the MFRM as well as the step difficulties in the LPCM.

3.4.3. Application to the Verbal Aggression data

Data

For the second empirical study, the verbal aggression data set (Vansteelandt, 2000) was used. This data set is publicly available at the BEAR center website page for De Boeck and Wilson (2004). (The data set can be downloaded from http://bearcenter.berkeley.edu/EIRM/). Participants were first-year psychology students of a Dutch speaking Belgian university. 24 items were presented in Dutch, the native language of all the participants, to ask behavioral questions about verbally aggressive reactions to frustrating situations. A total of 316 persons responded to the 24 items so that there were 22,752 observations[9] without missing values.

The item responses were three ordered-category responses (no = 0, perhaps = 1, and yes = 2). This data set has been diversely analyzed in measurement modeling contexts such as in Choi and Wilson (2015), De Boeck and Wilson (2004), and De Boeck (2008), and many of those analyses have been done after a dichotomization of the ordered-category responses. In this empirical study, the original ordered-category responses were used as polytomous data.

All items were designed to have a stem which describes a frustrating situation and a verbal aggression response part which describes how people could respond to the situation in question. In this item design, there were three experimental design factors[10]:

(a) the Behavior Mode factor has two levels of modes—wanting (Want) and doing (Do),
(b) the Situation Type factor has two types of situations—situations in which someone else is to blame (Other-to-blame) and situations in which oneself is to blame (Self-to-blame),
(c) the Behavior Type factor has three kinds of verbal aggressive behaviors which represent the extent to which they ascribe blame and the extent to which they express frustration—cursing (Curse), scolding (Scold), and shouting (Shout). Cursing is regarded as blaming and expressive, scolding is primarily blaming, and shouting is primarily expressive.

For the verbal aggression response, one of the two behavioral modes could be combined with one of the three verbal aggressive behaviors. For example, combinations of these two factors generate six responses such as "I would curse" (for doing and cursing) and "I would want to scold." (for wanting and scolding).

For the item stem, four frustrating situations were used:

Other-to-blame A. A bus fails to stop for me. (Bus)
Other-to-blame B. I miss a train because the clerk gave me faulty information. (Train)
Self-to-blame A. The grocery store closes just as I am about to enter. (Store)
Self-to-blame B. The operator disconnects me when I used up my last 10 cents for a call. (Call)

[9] When using the *gllamm* command, item response data are long-formatted and expanded over all category scores for each item. Since all items had three ordered-category responses, there were 22,752 (316 × 24 × 3) observations without missing values.
[10] The experimental design factors had originally four, but a design factor, the specific situations that are ask about, is nested within the second factor (De Boeck & Wilson, 2004). So, I used them as three item properties.

Table 3-7. The verbal aggression items composed of three item properties

Item		Behavior Mode		Situation Type		Behavior Type		
Num.	Situations	Want	Do	Other-to-blame	Self-to-blame	Curse	Scold	Shout
1	A bus fails to stop for me.	1	0	1	0	1	0	0
2		1	0	1	0	0	1	0
3		1	0	1	0	0	0	1
4	I miss a train because the clerk gave me faulty information.	1	0	1	0	1	0	0
5		1	0	1	0	0	1	0
6		1	0	1	0	0	0	1
7	The grocery store closes just as I am about to enter.	1	0	0	1	1	0	0
8		1	0	0	1	0	1	0
9		1	0	0	1	0	0	1
10	The operator disconnects me when I used up my last 10 cents for a call.	1	0	0	1	1	0	0
11		1	0	0	1	0	1	0
12		1	0	0	1	0	0	1
13	A bus fails to stop for me.	0	1	1	0	1	0	0
14		0	1	1	0	0	1	0
15		0	1	1	0	0	0	1
16	I miss a train because the clerk gave me faulty information.	0	1	1	0	1	0	0
17		0	1	1	0	0	1	0
18		0	1	1	0	0	0	1
19	The grocery store closes just as I am about to enter.	0	1	0	1	1	0	0
20		0	1	0	1	0	1	0
21		0	1	0	1	0	0	1
22	The operator disconnects me when I used up my last 10 cents for a call.	0	1	0	1	1	0	0
23		0	1	0	1	0	1	0
24		0	1	0	1	0	0	1

These four frustrating situations were designed in the second design factor: two other-to-blame situations and two self-to-blame situations. The full item contains one of the four frustrating situations for the item stem, and one of the six combinations of two behavioral modes and three behavior types for the verbal aggression response to the situation. After considering two specific situations of the same type (A and B) as replications, in total 24 items were written to fit a $2 \times 2 \times 3$ design with two replications within each cell. For example, the first item, "A bus fails to stop for me. I would want to curse.", and the fourth item, "I miss a train because the clerk gave me faulty information. I would want to curse", were designed and identified as a combination of the three factors of Want (Behavior Mode), Other-to-blame (Situation Type), and Curse (Behavior Type). The three item design factors are regarded as the item properties so that we can apply item explanatory models to this data set. With consideration for the item design factors, the verbal

aggression items can be classified by the predictors of each item property as in Table 3-7. These predictors are functioning as weights of the elementary components gathered into a Q matrix in the LLTM approach (see Kubinger, 2009; Poinstingl, 2009).

To incorporate the three categorical item properties in the design matrix, they were dummy coded so that the Want in the Behavior Mode property, the Self-to-blame in the Situation Type property, and the Shout in the Behavior Type property were a reference for each item property. The item properties would act as item specific properties or item-by-step specific properties based on their hypothesized effects on the item locations or the step difficulties. As in the first empirical study, both types of item properties were taken into account for applications of the two proposed polytomous item explanatory models to the verbal aggression data.

Results

The PCM and the two polytomous item explanatory models were fit to the verbal aggression data, according to the data analysis procedure. Table 3-8 shows the results of the fitted models: the MFRM variation (also called the MFRM here), the LPCM variation (also called the LPCM here), and the PCM.

The estimated person variance (σ_θ^2) was 0.88 for both polytomous item explanatory models and it was 0.93 in the PCM. Since standard item explanatory models such as the LLTM cannot explain the item parameters perfectly based on the observed item properties, the person variance should show shrinkage. Although the two models use different item explanatory approaches to polytomous items, the shrinkage effect of the person variance estimate was the same for both.

When comparing the goodness-of-fit, the LR test (compared to the PCM) was significant for the MFRM ($\chi^2(19) = 165.142$, $p < .001$) and the LPCM ($\chi^2(38) = 206.818$, $p < .001$), meaning that both the MFRM and the LPCM fit worse than the PCM, as expected. This result was confirmed by the AIC—the values for the two models were greater than the PCM (12737.47 for the PCM, 12864.61 for the MFRM, 12868.29 for the LPCM). This makes sense because the smaller number of estimated parameters in the item explanatory models made model fit (accuracy) much worse in spite of them being more parsimonious models.

However, this was not supported by the BIC: the value for the PCM was greater than the others (13131.06 for the PCM, 13105.58 for the MFRM, 12956.64 for the LPCM). Although this conflicted result doesn't look as expected, it is also understood reasonably with consideration for the difference between the AIC and BIC. They differ in theoretical motivations, objectives, assumptions, and meaning of penalty terms (Kuha, 2004; Wagenmakers & Farrell, 2004). Basically, the BIC penalizes complexity more heavily for the number of freely estimated parameters, whereas the AIC penalizes a model with more parameters much less than the BIC (Kuha, 2004). In other words, the BIC favors parsimonious models to a greater extent than the AIC does. In Table 3-8, the 48 parameters for item effects in the PCM were reduced to 29 in the MFRM and 10 in the LPCM. In terms of the BIC, we could say that the two polytomous item explanatory models performed better than the PCM in this empirical study. In addition, in the Bayesian aspect of the BIC, it assumes that the true data generating model is in the set of candidate models and it measures the degree of belief that a certain model is the true model which generates the observed data (Wagenmakers & Farrell, 2004). In this view, we could also say that the LPCM which had the lowest BIC was most likely to be the true data generating model for the verbal aggression data. Thus, this result implies that the three factorial design with two replications worked well for item

generation and the three item properties (design factors) had a high explanatory value for the verbal aggression items.

Comparing the LPCM to the MFRM, the LR test was significant ($\chi^2(19) = 41.68$, p = 0.002), which indicates that the MFRM fit better than the LPCM to the verbal aggression data. The lower AIC value for the MFRM supports this finding, which make sense because the larger number of the estimated parameters in the MFRM made model fit (accuracy) much better than the LPCM. However, the LPCM had the lower BIC value, meaning that the LPCM performed better because it was much more parsimonious than the MFRM. In this empirical study, we couldn't conclude that one was better than the other in terms of the goodness-of-fit. In fact, the two polytomous item explanatory models are different methodologically as well as practically.

Table 3-8. Performance of the MFRM and the LPCM on the verbal aggression data

		Partial Credit Model (PCM)	Many-Facet Rasch Model (MFRM)	Linear Partial Credit Model (LPCM)
Step difficulty				
Item 1	δ_{11}	-0.42	-0.73	-0.95
	δ_{12}	-0.08	-0.30	-0.13
Item 2	δ_{21}	0.13	0.14	-0.07
	δ_{22}	0.15	0.14	0.30
Item 3	δ_{31}	0.32	0.41	0.73
	δ_{32}	0.94	1.11	0.80
Item 4	δ_{41}	-0.97	-0.99	-0.95
	δ_{42}	-0.03	-0.05	-0.13
Item 5	δ_{51}	-0.03	0.03	-0.07
	δ_{52}	0.20	0.25	0.30
Item 6	δ_{61}	0.50	0.65	0.73
	δ_{62}	0.59	0.87	0.80
Item 7	δ_{71}	-0.13	-0.23	-0.25
	δ_{72}	1.00	0.82	0.84
Item 8	δ_{81}	0.82	0.66	0.63
	δ_{82}	1.87	1.24	1.27
Item 9	δ_{91}	1.49	1.40	1.42
	δ_{92}	2.48	1.75	1.77
Item 10	δ_{101}	-0.56	-0.32	-0.25
	δ_{102}	0.66	0.92	0.84
Item 11	δ_{111}	0.66	0.69	0.63
	δ_{112}	1.16	1.21	1.27
Item 12	δ_{121}	1.28	1.38	1.42
	δ_{122}	1.31	1.77	1.77
Item 13	δ_{131}	-0.53	-0.42	-0.41
	δ_{132}	0.18	0.25	0.23
Item 14	δ_{141}	0.14	0.32	0.47
	δ_{142}	0.56	0.82	0.66
Item 15	δ_{151}	1.15	1.17	1.27
	δ_{152}	1.19	1.23	1.16

Table 3-8(Continued). Performance of the MFRM and the LPCM on the verbal aggression data

		Partial Credit Model (PCM)	Many-Facet Rasch Model (MFRM)	Linear Partial Credit Model (LPCM)
Step difficulty				
Item 16	δ_{161}	-0.18	-0.26	-0.41
	δ_{162}	0.18	0.09	0.23
Item 17	δ_{171}	0.46	0.40	0.47
	δ_{172}	0.90	0.74	0.66
Item 18	δ_{181}	1.62	1.49	1.27
	δ_{182}	1.55	0.90	1.16
Item 19	δ_{191}	0.41	0.26	0.28
	δ_{192}	1.68	1.20	1.19
Item 20	δ_{201}	1.51	1.38	1.16
	δ_{202}	2.21	1.39	1.63
Item 21	δ_{211}	2.76	2.63	1.96
	δ_{212}	3.08	1.39	2.12
Item 22	δ_{221}	-0.23	0.13	0.28
	δ_{222}	0.75	1.34	1.19
Item 23	δ_{231}	0.65	0.82	1.16
	δ_{232}	1.33	1.94	1.63
Item 24	δ_{241}	1.99	1.99	1.96
	δ_{242}	2.07	2.03	2.12
Variance				
	σ_θ^2	0.93	0.88	0.88
Goodness-of-fit				
	D	12639.47	12804.61	12846.29
	AIC	12737.47	12864.61	12868.29
	BIC	13131.06	13105.58	12956.64
	k	49	30	11

Note: k = The number of estimated parameters.

In addition, a graphical comparison was conducted to examine agreement for the estimated and calculated step difficulties between the fitted models. The constructed step difficulties δ'_{im} which were calculated from the estimated item property effects in each item explanatory model were compared to the directly estimated step difficulties δ_{im} for each item in the PCM. Figure 3-3 shows similar graphical agreement between the two item explanatory models. Compared to the LPCM, the MFRM had more step difficulty points close to the 45-degree line indicating perfect alignment but a few points were farther from the line. The second step difficulty of item 21 were located farthest from the 45-degree line in both models.

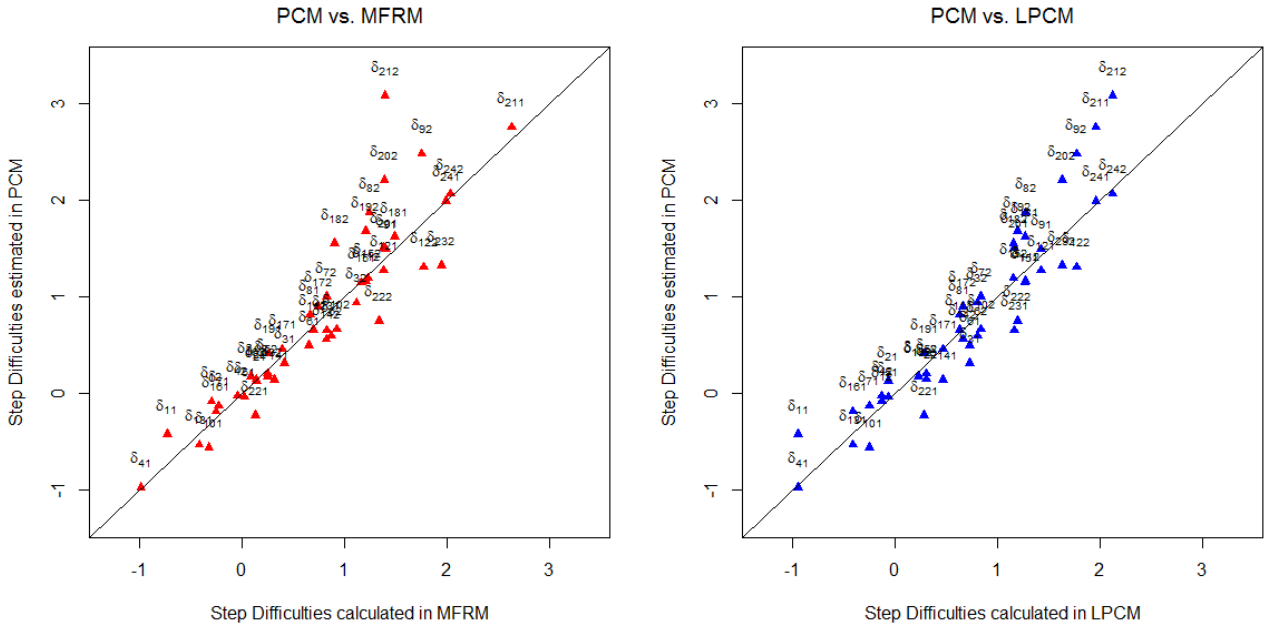


Figure 3-3. Graphical comparison of step difficulties between the models fitted to the verbal aggression data

In Table 3-9, a correlation analysis shows the LPCM had a slightly higher correlation with the PCM than the MFRM (ρ = 0.909 for the MFRM, and ρ = 0.930 for the LPCM), which means that the LPCM could predict the PCM step difficulties slightly better than the MFRM. A correlation of the calculated step difficulties between the two item explanatory models (ρ = 0.965) was higher than their correlations with the estimated step difficulties in the PCM, but it was not the perfect agreement. Considering the results from the graphical comparison, the two item explanatory models showed comparable performance in reconstructing the step difficulties using the item property effects.

Table 3-9. Correlations of the estimated and calculated step difficulties in the verbal aggression items

	Partial Credit Model (PCM)	Many-Facet Rasch Model (MFRM)	Linear Partial Credit Model (LPCM)
PCM	1	0.909	0.930
MFRM		1	0.965

The effects of the three item properties on the overall item difficulties or the step difficulties were reported separately for each model to show practical difference between the two polytomous item explanatory models. Recall that the verbal aggression data have three categorical item properties (Behavior Mode, Situation Type, and Behavior Type) from the item design factors. For

the MFRM, the three item properties incorporated as item specific properties were taken into account to explain and predict the overall item difficulties (item locations). Table 3-10 represents the results of the item property effects in the MFRM variation. In this model, the item specific property effect parameters γ are our focus to interpret the results (but the step deviation parameters τ are not of interest for interpretation). All predictors of the item property effects including the intercept were statistically significant.

Table 3-10. Item property effects on the verbal aggression items in the MFRM variation

Predictor	Item Parameter	Estimate	Standard Error	P-value
Intercept	γ_0	1.58	0.07	< .001
Behavior Mode (Do)	γ_{Do}	0.43	0.04	< .001
Situation Type (Other-to-blame)	γ_{Other}	-0.81	0.04	< .001
Behavior Type (Curse)	γ_{Curse}	-1.28	0.05	< .001
Behavior Type (Scold)	γ_{Scold}	-0.63	0.05	< .001
Step deviation parameters for each item	τ_{11}	-0.22	0.13	0.085
	τ_{21}	0.00	0.13	0.995
	τ_{31}	-0.35	0.13	0.005
	τ_{41}	-0.47	0.12	< .001
	τ_{51}	-0.11	0.13	0.377
	τ_{61}	-0.11	0.13	0.405
	τ_{71}	-0.52	0.12	< .001
	τ_{81}	-0.29	0.13	0.025
	τ_{91}	-0.18	0.15	0.242
	τ_{101}	-0.62	0.12	< .001
	τ_{111}	-0.26	0.13	0.048
	τ_{121}	-0.20	0.15	0.188
	τ_{131}	-0.34	0.12	0.006
	τ_{141}	-0.25	0.13	0.045
	τ_{151}	-0.03	0.14	0.833
	τ_{161}	-0.17	0.12	0.165
	τ_{171}	-0.17	0.13	0.173
	τ_{181}	0.29	0.16	0.062
	τ_{191}	-0.47	0.12	< .001
	τ_{201}	0.00	0.15	0.991
	τ_{211}	0.62	0.22	0.004
	τ_{221}	-0.60	0.12	< .001
	τ_{231}	-0.56	0.13	< .001
	τ_{241}	-0.02	0.17	0.911

For the Behavior Mode, holding other properties constant, the Doing mode made the overall item difficulty of an item 0.43 logits more difficult than the Wanting mode (γ_{Do} = 0.43, p < .001). In the Situation Type, the Other-to-blame situation made the overall item difficulty of an item 0.81

logits easier than the Self-to-blame situation (γ_{Other} = -0.81, p < .001), after keeping other properties constant. For the Behavior Type, holding other properties constant, compared to Shouting, Cursing made the overall item difficulty of an item 1.28 logits easier (γ_{Curse} = -1.28, p < .001), and Scolding made the overall item difficulty of an item 0.63 logits easier (γ_{Scold} = -0.63, p < .001).

In addition, 24 step deviation parameters were estimated for each item. By using the estimated item property effects and step deviation parameters, the constructed step difficulties for individual items could be calculated. For example, the first step difficulty of the first item ("A bus fails to stop for me. I would want to curse."), δ'_{11} was calculated as 1.58 (Intercept; γ_0) + 0.00 (Want; reference) − 0.81 (Other-to-blame; γ_{Other}) − 1.28 (Curse; γ_{Curse}) − 0.22 (τ_{11}) = − 0.73.

Table 3-11 shows the results of the item property effects in the LPCM variation. In this model, the three item properties incorporated as item-by-step specific properties were taken into account to explain and predict the step difficulties for each step. The estimated item-by-step specific property effect parameters ω were examined and interpreted to see how the three item properties affect the step difficulties step-specifically. For the verbal aggression items, the step difficulties for the first step represent the relative difficulties as one answers "perhaps" rather than "no", and those for the second step means the relative difficulties as one answers "yes" rather than "perhaps". All predictors of the item property effects including the step intercepts were statistically significant.

Table 3-11. Item property effects on the verbal aggression items in the LPCM variation

Predictor	Item Parameter	Estimate	Standard Error	P-value
Intercept	ω_{01}	1.42	0.09	< .001
	ω_{02}	1.77	0.12	< .001
Behavior Mode (Do)	ω_{Do1}	0.54	0.06	< .001
	ω_{Do2}	0.35	0.07	< .001
Situation Type (Other-to-blame)	ω_{Other1}	-0.69	0.06	< .001
	ω_{Other2}	-0.97	0.08	< .001
Behavior Type (Curse)	ω_{Curse1}	-1.68	0.08	< .001
	ω_{Curse2}	-0.93	0.10	< .001
Behavior Type (Scold)	ω_{Scold1}	-0.80	0.07	< .001
	ω_{Scold2}	-0.50	0.10	< .001

The effects of the three item properties on the step difficulties of the first step were interpreted as follows. When a person answered "perhaps" rather than "no", keeping other properties constant, the Doing mode made the items 0.54 logits more difficult than the Wanting mode (ω_{Do1} = 0.54, p < .001), the Other-to-blame situation made the items 0.69 logits easier than the Self-to-blame situation (ω_{Other1} = -0.69, p < .001), and compared to Shouting, Cursing made

the items 1.68 logits easier (ω_{Curse1} = -1.68, p < .001) and Scolding made the items 0.80 logits easier (ω_{Scold1} = -0.80, p < .001). For the step difficulties of the second step, holding other properties constant, the Doing mode made the items 0.35 logits more difficult than the Wanting mode (ω_{Do2} = 0.35, p < .001), the Other-to-blame situation made the items 0.97 logits easier than the Self-to-blame situation (ω_{Other2} = -0.97, p < .001), and compared to Shouting, Cursing made the items 0.93 logits easier (ω_{Curse2} = -0.93, p < .001) and Scolding made the items 0.50 logits easier (ω_{Scold2} = -0.50, p < .001), as a person answered "yes" rather than "perhaps". These results nicely agree with those obtained by Tuerlinckx and Wang (2004) except for the step intercept values, because they also added person predictors.

In order to reconstruct the step difficulties for individual items, we can calculate them by weighted sums of the estimated item property effects. For example, the first step difficulty of the first item ("A bus fails to stop for me. I would want to curse."), δ'_{11} was calculated as 1.42 (Intercept; ω_{01}) + 0.00 (Want; reference) − 0.69 (Other-to-blame; ω_{Other1}) − 1.68 (Curse; ω_{Curse1}) = − 0.95.

In sum, there was no uniform agreement in performance of the MFRM and the LPCM in terms of the goodness-of-fit. The comparable performance of the two polytomous item explanatory models was found in reconstructing the step difficulties based on the estimated item property effects. The two different item explanatory approaches showed practical differences in interpreting the effects of the item properties on the verbal aggression items.

3.5. Conclusion and Discussion

I have investigated how to apply the LLTM approach, namely item explanatory modeling, to polytomous data under the MGLMM framework, given the ordered-category responses and the adjacent-categories logit link. The two item explanatory extensions of the PCM, the MFRM variation and the LPCM variation, were discussed and specified with formulations. These two polytomous item explanatory IRT models were fit to the Carbon Cycle assessment data and to the verbal aggression data respectively in the empirical studies.

The first empirical study showed that the MFRM variation had a superior goodness-of-fit than the LPCM variation in terms of both the AIC and BIC, whereas in the second empirical study there was no uniform agreement between the two goodness-of-fit indices to see performance of the MFRM variation and the LPCM variation. In both studies, the two polytomous item explanatory models showed comparable performance in reconstructing the step difficulties using the estimated item property effects but they demonstrated practical differences in interpreting the effects of the item properties on the polytomous item difficulties. The MFRM variation could explain and predict the overall item difficulties (item locations) by the item properties incorporated as item specific properties, and the LPCM variation could explain and predict the step difficulties by the item properties incorporated as item-by-step specific properties. Thus, the MFRM variation and the LPCM variation worked well for the two empirical data sets in a similar but different way, as they both gave us informative feedback for enhancing quality of item development in practice.

In fact, the two polytomous item explanatory models are methodologically and practically different in terms of the target difficulty parameters of polytomous items which are explained by item properties, the types of item properties incorporated in the design matrix, and the types of

item property effects. Before fitting the model to the data, it is highly recommended for model selection to clarify which polytomous item difficulties are explained by item properties as well as to examine what kind of item properties are included in the design matrix. When designing items, it is helpful to examine underlying hypotheses for item properties or design factors. If an item property is assumed to affect specific levels of the construct as in the learning progression framework (see Table 3-1), the LPCM variation is useful to test that hypothesis. If one assumes that the item property has nothing to do with the specific levels, the MFRM variation is appropriate to see its influence on the overall construct. One recommendable and systematic way of designing items is to use Wilson's (2005) four building blocks approach to constructing measures, which takes advantage of the principles of sound educational measurement. By using an iterative cycle of the measurement processes—construct mapping, item design, outcome space, and measurement model, one can develop items with articulating the underlying hypotheses.

Among the EIRM approaches, I have focused on explaining the item side only of the response data. There are huge potentials to extend the polytomous item explanatory IRT models. For instance, when we consider to add person predictors and/or person-by-item predictors, they can be extended with other explanatory models such as a latent regression and differential item functioning analysis. A future study can address multidimensional extensions of those models to examine interactions between item properties and persons, which are applications of the Random-Weights Linear Logistic Test Model (RW-LLTM; Rijmen & De Boeck, 2002) to polytomous data. In addition, when we are not sure of perfect explanation through the observed item properties, random item errors can be added to the polytomous item explanatory models, which are applications of the Linear Logistic Test Model with random item error (LLTM + ε; Mislevy, 1988; Janssen, Schepers, & Peres, 2004) to polytomous data.

This paper also shed light on the methodological advantages of item explanatory modeling. In the polytomous item explanatory models I have specified in this paper, the step difficulty parameters or the overall item difficulty parameters are decomposed into weighted sums of the item property effect parameters. These models enable us to estimate a smaller number of item parameters than the saturated model to explain and predict the item effects, which is a methodological advantage in extracting essential and meaningful elementary components by incorporating item properties. It is useful to understand how the items generated by item design or test construction work, as well as to validate hypothesized constructs for item design or test construction. This can also help us generate new items by composing specified combinations of the item property effects so that we can predict item locations or step difficulties of the newly developed items in a more scientific and systematic manner rather than an intuitive manner. It is promising that the polytomous item explanatory models have been widely used to measure the effects of various testing conditions such as raters, cognitive operations, and item exposure time, as the many have done for dichotomous data (e.g., Cristante & Robusto, 2007; Draney & Wilson, 2008; Fischer, 1973; Kubinger, 2009; Lei, 2009; Poinstingl, 2009; Sonnleitner, 2008).

In conclusion, we can say that the polytomous item explanatory IRT models can have a large effect in methodological foundations for the educational measurement, particularly with consideration for De Boeck and Wilson's (2004) EIRM framework. Peter Drucker, a famous management thinker, said that *"If we can't measure it, we can't manage and improve it."* Likewise, if we cannot explain the measurement, we cannot know how to improve it.

Chapter 4.
Polytomous Item Explanatory IRT Models with Random Item Effects

4.1. Introduction

The primary role of educational measurement and assessment is to provide necessary information about ways to facilitate teachers' instruction and students' learning (Black & Wiliam, 1998). This in turn emphasizes the capability and quality of informative measurement and assessment. Explanatory measurement (De Boeck & Wilson, 2004) provides various explanatory inferences from the assessments so that it can strengthen the kind of feedback that could be given to both teachers and students as well as to test developers and educational researchers. In item response theory (IRT), explanatory item response models (EIRM; De Boeck & Wilson, 2004) aim to explain the person and/or item side of the assessment data in order to enrich inferential information and enhance the feedback. Among person explanatory, item explanatory, and doubly explanatory models of the EIRM approach, this paper will focus on item explanatory models in which item properties are incorporated to explain and predict the item effects.

A typical approach to item explanatory modeling is the Linear Logistic Test Model (LLTM; Fischer, 1973), which decomposes the difficulties of specific items into linear combinations of elementary components related to item properties or features (Embretson & Reise, 2000). In measurement practices, this LLTM approach has various methodological advantages in extracting essential and meaningful elementary components, testing constructs hypothesized in item design and generation, and predicting item difficulties of newly developing items as well as in measuring the effect of various testing conditions such as item presentation position and item exposure time (Cristante & Robusto, 2007; Embretson & Gorin, 2001; Kubinger, 2009; Poinstingl, 2009). This approach is also useful to examine the effect of item properties such as item design variables, content-specific learning, and cognitive operations in various assessment contexts. Thus, item explanatory models using the LLTM approach can serve to provide useful and practical information for enhancing test development, item generation, and educational assessment.

Nevertheless, the original approach of LLTM has an underlying assumption that predictors of the observed item properties or elementary components can "perfectly" account for item difficulties (De Boeck, 2008; De Boeck & Wilson, 2004). However, "perfect" explanation is hardly possible because substantive theories behind the measurement model may not be flawless and/or the item difficulty parameter may be a random variable by nature (De Boeck, 2008; Janssen, Schepers, & Peres, 2004). Considering the uncertainty in explanation and/or the random nature of item parameters, as in an ordinary regression model, it is reasonable to add a random error or residual term into the item regression component of item explanatory models. This approach is the Linear Logistic Test Model with item error (LLTM + ε; Mislevy, 1988; Janssen, Schepers, & Peres, 2004), which enhances prediction of the item difficulty parameters of a Rasch model by allowing for residual variation (De Boeck, 2008; Hartig, Frey, Nold, & Klieme, 2012). That is, an item error term in the LLTM + ε can compensate for the discrepancy between the freely estimated item difficulty in the Rasch model and the item difficulty calculated by the estimated item property effects in the LLTM. In particular, the LLTM + ε is a type of random item IRT model (De Boeck,

2008). The LLTM assumes that the item difficulties are perfectly predicted by the fixed item property effects, whereas the LLTM + ε relaxes this assumption by allowing for random variation of each item. Random item IRT models is a rather new area in educational and psychological measurement research (De Boeck, 2008). It is a usual the practice in IRT to deal with items or item covariates as fixed and persons as random, however, in random item IRT models, items are treated as random and item difficulties are regarded as random effects and hence the item difficulty parameter is a random variable.

In educational assessment contexts, students' achievement outcomes are typically item responses which are observed as categorical data, either dichotomous (binary) or polytomous (multi-categorical). In this paper, we mainly consider ordered-category responses, because it is common to have ordered-category responses in a wide range of educational, psychological, and sociological applications (De Boeck & Wilson, 2004). In educational assessment practices, for instance, a learning progression is defined as hypothetical descriptions of the successively more sophisticated levels of students' understanding and learning about a specific topic over time (Duncan & Hmelo-Silver, 2009). A simple learning progression can be conceptualized and measured through a construct map (Wilson, 2009). Based on the construct map, the levels of student achievement are scored and provided in the form of ordered-category responses (e.g. Jin & Anderson, 2012; Jolin, 2015).

For polytomous extensions of item explanatory models, referred to as polytomous item explanatory models, it is appealing to apply the LLTM + ε approach to polytomous data with consideration for the uncertainty in explanation and/or the random nature of item parameters. However, polytomous extensions are more complicated than applications to dichotomous data. This originates from mainly two issues: a modeling issue and an estimation issue. First, re-parameterizing polytomous item parameters by incorporating item properties is complicated in polytomous item explanatory models, and adding different types of random item errors into those models is also difficult. Second, treating both items and persons as random makes for crossed random effects (De Boeck, 2008; Janssen, Schepers, & Peres, 2004). Estimation of random item IRT models with the crossed random effects is demanding due to the complexity and difficulty in numerical integration (Cho & Rabe-Hesketh; 2011; Van den Noortgate, De Boeck, & Meulders, 2003), which becomes exacerbated and practically difficult for polytomous data. These issues may be why few studies have investigated polytomous extensions of the LLTM + ε approach. Most of its applications have examined dichotomous data rather than polytomous data (e.g. De Boeck, 2008; De Boeck et al., 2011; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009).

This research aims to develop polytomous item explanatory models by extending the LLTM + ε approach to polytomous data. To resolve the modeling issue against its polytomous extensions, two steps of measurement modeling are required. First, I will address random item effects in polytomous item response models in terms of a random sampling interpretation. This enables one figure out the underlying distribution of random item errors on the polytomous item parameters. Second, I will examine polytomous item explanatory models and then add random item errors to those models in terms of a random error interpretation. In addition, to resolve the estimation issue, I will review available estimation methods for the crossed random effects models and then discuss estimation of the proposed polytomous item explanatory models with random item errors using a Bayesian inference approach. Furthermore, to see how the proposed models work in different explanatory conditions, two simulation studies will be conducted to assess model comparison and parameter recovery as well as effects of model misspecification. Lastly, two empirical studies

focusing on practical implications and applications of those models to the Carbon Cycle assessment data and the verbal aggression data will be demonstrated.

4.2. Polytomous Item Response Models with Random Item Effects

4.2.1. Rasch Family Models for Polytomous data

This section reviews polytomous item response models first to investigate random item effects for those models. Given a context of ordered-category responses regarded as polytomous data, adjacent-categories logits will be employed for polytomous item response models and their item explanatory extensions. Since most of the ordered-category responses in educational assessment or cognitive development contexts are subjectively assigned scores between categories, adjacent-categories logits are suitable for modeling those scores in ordinal regression models (Anderson, 1984; Johnson, 2007). Moreover, adjacent-categories logits are useful when contrasting probabilities of the responses in pairs of adjacent categories (Hartzel, Agresti, & Caffo, 2001; Mellenbergh, 1995), regarded as the local comparison of ordered categories (Masters & Wright, 1997).

Rasch family models for polytomous data, which are based on the adjacent-categories logits, can make person and item parameters separable and hence have sufficient statistics as well as make possible specifically objective comparisons of persons and items (Masters, 1982; Masters & Wright, 1997). These are fundamental measurement properties of the Rasch family models (Masters & Wright, 1984; Rost, 2001). In particular, the Partial Credit Model (PCM; Masters, 1982) is a straightforward application of the Rasch model (Rasch, 1960) originally for dichotomous responses to pairs of adjacent categories in a sequence in ordered-category responses (Masters & Wright, 1997). Hence, I will use the PCM as a basic polytomous item response model to investigate random item effects in polytomous items and add random item errors to polytomous item explanatory models.

To facilitate interpretation of polytomous item parameters, explanatory convenience, and simplicity of model specification, the PCM is expressed in terms of the local comparison between adjacent categories. A comparison of the response probabilities that person p of ability θ_p scores from $m-1$ to m on item i ($i = 1,2,\dots,\mathrm{I}$), which is regarded as the m-th adjacent-categories logit (Tuerlinckx & Wang, 2004), can be written as:

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)} = \theta_p - \delta_{im}\,, \qquad m = 1,\dots,M_i\,, \tag{4.1}$$

where $\theta_p \sim N(0,\sigma_\theta^2)$ and $\delta_{i0} = 0$. For statistical model identification[11], the mean person ability is constrained to zero. From this local comparison perspective of the PCM, the item parameter δ_{im} is interpreted as a "step" difficulty when shifting a category score from $m-1$ to m on item i,

[11] For model identification, a particular item parameter or the mean of item parameters can be constrained to zero, instead of constraining the mean person ability (De Boeck & Wilson, 2004).

which is regarded as the m-th step of switching over to the next response category. Thus, a step difficulty parameter δ_{im} indicates the relative difficulty for scoring m rather than $m-1$ on item i (Masters, 1982; Masters & Wright, 1984). Note that in this formula, the person abilities θ_p are treated as random and the step difficulties δ_{im} are treated as fixed.

In particular, the step difficulty parameter δ_{im} can be split into two item parameters, the item location (a.k.a. overall item difficulty) parameter β_i and the step deviation parameter τ_{im}, as follows:

$$\delta_{im} = \beta_i + \tau_{im}, \tag{4.2}$$

so that

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)} = \theta_p - \beta_i - \tau_{im}\,, \qquad m=1,\dots,M_i\,, \tag{4.3}$$

where $\theta_p \sim N(0,\sigma_\theta^2)$, $\tau_{i0}=0$, and $\sum_{m=1}^{M_i}\tau_{im}=0$ so that $\frac{1}{M_i}\sum_{m=1}^{M_i}\delta_{im}=\beta_i$. Due to the model constraint that $\frac{1}{M_i}\sum_{m=1}^{M_i}\delta_{im}=\beta_i$, an item location parameter β_i is interpreted as the overall item difficulty for polytomous item i, and a step deviation parameter τ_{im} is interpreted as a deviation of the step difficulty from the overall item difficulty for the m-th step within item i ($\tau_{im}=\delta_{im}-\beta_i$). M_i-1 number of step deviation parameters are estimated for each item because the sum of step deviations is constrained to zero for each item ($\sum_{m=1}^{M_i}\tau_{im}=0$). This twofold item parameterization is used in the IRT software ConQuest (Adams, Wu, & Wilson, 2012), which is essential to identify overall item difficulties, scale thresholds, and step difficulties in polytomous items. It is also useful to incorporate item properties and add an error term in order to explain the overall item difficulties.

In addition, the Rating Scale Model (RSM; Andrich, 1978) is a special case of the PCM (Masters, 1982), in which the relative difficulties of the steps between categories do not vary across all items ($\tau_{im}=\tau_m$) and the number of steps are the same for all items ($M_i=M$). In the RSM, the item location parameter β_i and the common scale threshold parameter τ_m can be formulated as:

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)} = \theta_p - \beta_i - \tau_m\,, \qquad m=1,\dots,M\,, \tag{4.4}$$

where $\theta_p \sim N(0,\sigma_\theta^2)$, $\tau_0=0$, and $\sum_{m=1}^{M}\tau_m=0$. Note that the RSM imposes more restrictions on the step deviation parameter of the PCM. In sum, the PCM is the most general Rasch family model for polytomous (ordinal) data (Rost, 2001).

4.2.2. Polytomous Rasch Models with Random Item Effects

Based on the polytomous Rasch models, I will examine random item effects that account for the random nature of polytomous item parameters. In random item IRT models, random item effects can be interpreted in two ways (Janssen, Schepers, & Peres, 2004): a random error

interpretation and a random sampling interpretation. For example, in the LLTM + ε, the difficulties of specific items constructed by linear combinations of item property effects can be approximately true but not perfectly true. There can be residuals unexplained by item properties. Referring to a random error interpretation, this represents the uncertainty in explanation. In a random sampling interpretation, on the other hand, the difficulties for newly developing items can be predicted under the assumption that items are randomly sampled from a large item population and they share the same values on the linear combinations of item predictors. This suggests the random nature of item parameters. In fact, these two interpretations are two sides of the same coin because item difficulties are seen as random effects and their parameters are random variables in the model in both interpretations.

The notion of random item effects becomes clearer in the context of item banks, item generation (or item cloning), and automatic item generation (AIG) as well as generalizability in item response modeling (GIRM). In practice, there are some preexisting item populations. For instance, with the advent of computer adaptive testing (CAT), item banks or item pools are created using IRT models and they are used as the item population for random item selection (De Boeck, 2008; Glas & van der Linden, 2003; Haley et al., 2006). In CAT, items are randomly drawn from an item bank so that a homogenous set of the items can be constructed. Also, item generation is seen as formally equal to drawing from a theoretical population such as in a theory of domain-referenced testing (De Boeck, 2008; Hively, 1974). The concept of "domain", also referred to as "universe", means the knowledge and skills required for the mastery of a specific content area in educational assessment (Janssen, Schepers, & Peres, 2004). In AIG, items (or item clones) within an item family are generated automatically based on the item generating models and are administered on a computer-based test (Holling, Bertling, & Zeuch, 2009; Johnson & Sinharay, 2005; Lathrop & Cheng, 2017). The concept of 'item family', based on the principal of item generation, represents a set of items with sufficient communalities as well as sufficient differentiation from other sets of items (De Boeck, 2008; Glas & van der Linden, 2003; Sinharay, Johnson, & Williamson, 2003). Such an item universe or item family is regarded as the item population from which test items are randomly sampled or automatically generated. This implies the items are seen as random. Moreover, the random nature of the items concerns the generalizability potential of a random item effects model (De Boeck, 2008). As in the GIRM approach (Briggs & Wilson, 2007; Choi, 2013), generalization to the item universe can be a matter of concern even when the items are not randomly sampled in reality from a preexisting item population. Thus, random item effects models are promising and useful to calibrate item parameters from the underlying item population in various measurement situations.

Despite the practical implications and the potential usefulness in educational measurement practices, random item effects models are a rather new approach in IRT (De Boeck, 2008). In fact, little is known about random effect polytomous items. Most relevant studies that have investigated random item effects have used dichotomous data (e.g. Cho & Rabe-Hesketh; 2011; Cho et al., 2014; De Boeck, 2008; De Boeck et al., 2011; Glas & van der Linden, 2003; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009), whereas a few studies have discussed item selection techniques or item family calibration methods for polytomous items in the context of CAT and AIG (e.g. Johnson & Sinharay, 2005; Pastor, Dodd, & Chang, 2002; Penfield, 2006). It has been rare to investigate random item effects for polytomous item response models.

In the random item effects model setting, items are treated as random. The randomness implies that there could, in principle, be many items. In the item response process for polytomous

(ordinal) items, the scale structure is fixed by a scale type and outcome space corresponding to constructs for item design such as a learning progression; thus, the number of response categories is fixed and hence that of category steps is fixed. In order to come up with polytomous random item effects in terms of the random sampling interpretation, this paper focuses on the item parameters of the PCM, the most general Rasch family model for polytomous data. To allow for random item effects in the PCM, it is helpful to examine the conceptualization of random effects on the item side of the polytomous responses by considering different types of item parameters. To my knowledge, there are three types of random item effects: (a) overall item random effects—random variation in overall item difficulties across items, (b) step specific random item effects—correlated random variations in step difficulties between category steps across items, and (c) item-step random effects—independent (or uncorrelated) random variations in step difficulties across items and category steps.

For the first type of polytomous random item effects, by using the twofold item parameterization for the PCM in Equations 4.2 and 4.3, random effects can be imposed on the overall item difficulties for individual item so that the item location parameters β_i are a random variable, expressed as:

$$ln\frac{Pr(y_{pi}=m|\theta_p,\beta_i)}{Pr(y_{pi}=m-1|\theta_p,\beta_i)}=\theta_p-\beta_i-\tau_{im} \quad \text{and} \quad \beta_i\sim N(\mu_\beta,\,\sigma_\beta^2), \tag{4.5}$$

where $\theta_p \sim N(0,\sigma_\theta^2)$, $\tau_{i0}=0$, and $\sum_{m=1}^{M_i}\tau_{im}=0$. The overall item random effects, denoted by β_i, are assumed to follow a normal distribution with a mean of μ_β and a variance of σ_β^2. Here, μ_β represents a mean overall item difficulty across all items in the ordinal item population. For model identification, the mean person ability is constrained to zero. In this approach the sum of step deviations is constrained to zero for each item ($\sum_{m=1}^{M_i}\tau_{im}=0$) to identify the overall difficulty of an item, and the step deviation parameter τ_{im} is not random. In this approach, the explanation and prediction of the overall item difficulties are a matter of interest in interpretation of random item effects.

In the second type of polytomous random item effects, the explanation and prediction of the step difficulties are important. Given the same number of response categories and category steps from the same scale structure in the same item population, the step difficulties can vary within an item and between items as the items vary. In the PCM, the step difficulties of an item do not rely on the particulars of the other items in the sample from the item population and there are no order restrictions between them within an item (Curtis, 2010; Masters, 1982; Masters & Wright, 1997). However, within an item, the step difficulties are not interpreted independently each other (Masters & Wright, 1984; Molenaar, 1983). That is, category scores have some deterministic dependence on the relative difficulties between category steps in the same item despite the assumption of local independence between items given the person ability (Masters, 1982; Molenaar, 1997). This implies a distribution assumption for the ordinal item population that the step difficulties in each step can randomly vary across items but they are correlated within an item.

To accommodate these concerns, by using the original PCM's item parameterization in Equation 4.1, correlated item-step random effects between category steps can be imposed on the step difficulties step-specifically for individual item, referred to as step specific random item

effects, so that the step difficulty parameter vector $\boldsymbol{\delta_i}$ are a set of correlated random variables, as follows:

$$ln\frac{Pr(y_{pi}=m|\theta_p,\boldsymbol{\delta_i})}{Pr(y_{pi}=m-1|\theta_p,\boldsymbol{\delta_i})}=\theta_p-\delta_{im} \quad \text{and} \quad \boldsymbol{\delta_i}\,(\delta_{i1},\delta_{i2},\dots,\delta_{im})'\sim MVN_m(\boldsymbol{\mu_\delta},\boldsymbol{\Sigma}), \qquad (4.6)$$

where

$$\boldsymbol{\mu_\delta}=\begin{bmatrix}\mu_{\delta1}\\ \mu_{\delta2}\\ \vdots\\ \mu_{\delta m}\end{bmatrix},\ \boldsymbol{\Sigma}=\begin{bmatrix}\sigma^2_{\delta1} & \sigma_{\delta1\delta2} & \cdots & \sigma_{\delta1\delta m}\\ \sigma_{\delta2\delta1} & \sigma^2_{\delta2} & \cdots & \sigma_{\delta2\delta m}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{\delta m\delta1} & \sigma_{\delta m\delta2} & \cdots & \sigma^2_{\delta m}\end{bmatrix},\ \theta_p\sim N(0,\sigma^2_\theta), \text{and } \delta_{i0}=0.$$

The vector of the step specific random item effects $\boldsymbol{\delta_i}\,(\delta_{i1},\delta_{i2},\dots,\delta_{im})'$ is assumed to follow a multivariate normal (MVN) distribution with a mean vector $\boldsymbol{\mu_\delta}\,(\mu_{\delta1},\,\mu_{\delta2},\dots,\mu_{\delta m})'$ and a variance-covariance matrix $\boldsymbol{\Sigma}\,(m\times m)$. Here, $\mu_{\delta m}$ represents a mean step difficulty for the m-th step across all items in the ordinal item population. For model identification, the mean person ability is constrained to zero.

The third type of polytomous random item effects are independent item-step random effects that the step difficulties can randomly vary across items and category steps. By using the original PCM's item parameterization in Equation 4.1, uncorrelated random effects between category steps can be imposed on the step difficulties for individual item so that the step difficulty parameters δ_{im} are a random variable, written as:

$$ln\frac{Pr(y_{pi}=m|\theta_p,\delta_{im})}{Pr(y_{pi}=m-1|\theta_p,\delta_{im})}=\theta_p-\delta_{im} \quad \text{and} \quad \delta_{im}\sim N(\mu_\delta,\,\sigma^2_\delta), \qquad (4.7)$$

where $\theta_p\sim N(0,\sigma^2_\theta)$ and $\delta_{i0}=0$. The item-step random effects δ_{im} is assumed to follow a normal distribution with a mean of μ_δ and a variance of σ^2_δ. Here, μ_δ represents a grand mean step difficulty across all items and category steps in the ordinal item population. For model identification, the mean person ability is constrained to zero.

In fact, this item-step random effects model is a special case of the step specific random item effects model in Equation 4.6. The two models are identical with additional assumptions for the latter that the mean and variance of the step specific random item effects are the same across every step ($\mu_{\delta1}=\cdots=\mu_{\delta m}=\mu_\delta$ in $\boldsymbol{\mu_\delta}$; $\sigma^2_{\delta1}=\cdots=\sigma^2_{\delta m}=\sigma^2_\delta$ in $\boldsymbol{\Sigma}$) and the step specific random item effects are independent between category steps within an item (zero correlation or zero covariance; $\sigma_{\delta1\delta2}=\cdots=\sigma_{\delta m\delta m-1}=0$ in $\boldsymbol{\Sigma}$). The item-step random effects seem to be peculiar due to assumptions that do not reflect the multi-categorical scale structure in the ordinal items. Nevertheless, this type of polytomous random item effects has been commonly used for a prior distribution of the step difficulty parameters in Bayesian estimation methods (e.g. Ames & Samonte, 2015; Curtis, 2010; Fox, 2010; Kang, Cohen, & Sung, 2009). This makes sense in that a weakly or non-informative prior for model parameters need not be strictly true in Bayesian inference (Fox, 2010; Stan Development Team, 2017).

The three types of polytomous random item effects for the PCM are summarized with each distribution assumption in Table 4-1. Since the two interpretations of random item effects are closely related, a distribution assumption for each polytomous random item effects model will be retained for corresponding random item errors in the following sections.

Table 4-1. Three types of polytomous random item effects for the PCM

Item Parameterization	Random Item Parameter	Polytomous Random Item Effects	Distribution Assumption
Two fold $\delta_{im} = \beta_i + \tau_{im}$	Item Location Parameter β_i	Overall Item Random Effects	$\beta_i \sim N(\mu_\beta,\ \sigma_\beta^2)$ and $\sum_{m=1}^{M_i} \tau_{im} = 0$
One fold δ_{im}	Step Difficulty Parameter δ_{im}	Step Specific Random Item Effects	$\boldsymbol{\delta_i}\ (\delta_{i1}, \dots, \delta_{im})' \sim MVN_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
One fold δ_{im}	Step Difficulty Parameter δ_{im}	Item-Step Random Effects	$\delta_{im} \sim N(\mu_\delta,\ \sigma_\delta^2)$

Note. The overall item random effects β_i follow a normal distribution with a mean of μ_β and a variance of σ_β^2. The vector of the step specific random item effects $\boldsymbol{\delta_i}\ (\delta_{i1}, \delta_{i2}, \dots, \delta_{im})'$ follows a multivariate normal distribution with a mean vector $\boldsymbol{\mu_\delta}\ (\mu_{\delta 1},\ \mu_{\delta 2}, \dots, \mu_{\delta m})'$ and a variance-covariance matrix $\boldsymbol{\Sigma}\ (m \times m)$. The item-step random effects δ_{im} follow a normal distribution with a mean of μ_δ and a variance of σ_δ^2.

In addition, the step specific random item effects model in Equation 4.6 hints at a random item effects version of the RSM. As in the overall item random effects model in Equation 4.5, the item location parameters β_i can be a random variable which follows a normal distribution with a mean of μ_β and a variance of σ_β^2. Based on the RSM assumption, the relative difficulties between category steps are the same for all items so that the common scale threshold parameter τ_m in Equation 4.4 concerns only the rating scale structure. By summing up μ_β and τ_m, one can identify a similar mean vector with the step specific random item effects model. A further investigation for this topic is needed but I will leave it for a future study in order to focus on extensions of the PCM in this paper.

4.3. Polytomous Item Explanatory Models with Random Item Errors

4.3.1. Polytomous Item Explanatory Models

For the purpose of developing polytomous item explanatory models with random item errors, it is necessary to investigate item explanatory extensions of a polytomous item response model before adding random item errors. Since I set the PCM as a basic model for those incremental extensions, I will apply the LLTM approach to the PCM for item explanatory extensions first. However, re-parameterizing polytomous item parameters by incorporating item properties into the PCM is complicated. To avoid the complication of item parameterization, it is

helpful to clarify the target difficulty parameters in the PCM which are explained by item properties: (a) the item location (overall item difficulty) parameters β_i can be explained by item properties, and (b) the step difficulty parameters δ_{im} can be explained by item properties.

Many-Facet Rasch Model with a decomposition of the item location parameters

In the first case, by using the twofold item parameterization for the PCM in Equation 4.3, we can impose linear restrictions on the item location parameters as was done in the LLTM, and also estimate the step deviation parameters for each item. The restricted item location parameters β_i' are decomposed into weighted sums of item property effect parameters γ_k as follows:

$$\beta_i' = \sum_{k=0}^{K} \gamma_k x_{ik}\,, \qquad k = 0,\ \ldots,\ K \tag{4.8}$$

so that

$$ln\frac{Pr\left(y_{pi} = m\middle|\theta_p\right)}{Pr\left(y_{pi} = m-1\middle|\theta_p\right)} = \theta_p - \sum_{k=0}^{K} \gamma_k x_{ik} - \tau_{im}, \tag{4.9}$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$, $\tau_{i0} = 0$, and $\sum_{m=1}^{M_i} \tau_{im} = 0$. Also, where γ_0 is the item intercept representing the difficulty for items with all x_{ik} = 0 for k > 0, γ_k is the regression weight or the effect of item property k on the overall item difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, x_{ik} is the value of item i on item property k, and τ_{im} is a step deviation parameter for the m-th step of item i. For model identification, the mean person ability is constrained to zero. Note that in general the restricted item location parameters β_i' in Equation 4.8 will not be equal to the original parameters β_i in Equation 4.3 because the explanation by the observed item properties are not perfect.

This item explanatory extension of the PCM is a variation of Linacre (1989)'s Many-Facet Rasch Model (MFRM) in that the item location parameters are decomposed into a linear combination of the effects of item properties which are regarded as sub-facets within the item facet (Wang & Wilson, 2005). Building on the MFRM approach, more general regression notation is used to include both categorical and continuous explanatory predictors, and the step deviation parameters are estimated for each item. This polytomous item explanatory model will be called the "Many-Facet Rasch Model with a decomposition of the item location parameters".

Linear Partial Credit Model with a decomposition of the step difficulty parameters

For the second case, we can impose linear restrictions on the step difficulty parameters in Equation 4.1 by incorporating item properties. The restricted step difficulty parameters δ_{im}' are decomposed into weighted sums of step specific item property effect parameters ω_{km} as follows:

$$\delta_{im}' = \sum_{k=0}^{K} \omega_{km} x_{ik}\,, \qquad k = 0,\ \ldots,\ K \tag{4.10}$$

so that

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)}=\theta_p-\sum_{k=0}^{K}\omega_{km}x_{ik}\,, \tag{4.11}$$

where $\theta_p \sim N(0,\sigma_\theta^2)$, $\omega_{k0}=0$, and also where ω_{0m} is the step intercept representing the m-th step difficulty for items with all x_{ik} = 0 for k > 0, ω_{km} is the regression weight or the effect of item property k on the m-th step difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, and x_{ik} is the value of item i on item property k. For model identification, the mean person ability is constrained to zero. Note that in general the restricted step difficulty parameters δ'_{im} in Equation 4.10 will not equal the original parameters δ_{im} in Equation 4.1 because there is no complete explanation via the observed item properties.

This item explanatory extension of the PCM is a variation of the Linear Partial Credit Model (LPCM; Fischer and Ponocny, 1994) in that the step difficulty parameters are decomposed into a linear combination of item properties, although they used complicated and less interpretable item parameters such as a normalization constant and basic parameter. Building on the LPCM approach, the step specific item property effect parameters are used to interpret the effects of item properties on the step difficulties for each step. This polytomous item explanatory model will be called the "Linear Partial Credit Model with a decomposition of the step difficulty parameters".

4.3.2. Polytomous Item Explanatory Models with Random Item Errors

As the last step, in terms of a random error interpretation of random item effects, I will add random item errors to the polytomous item explanatory models. The polytomous random item effects have an underlying distribution which can be used for each corresponding type of polytomous random item error. As in the LLTM + ε approach, an error term will enhance prediction of the polytomous item difficulties, the overall item difficulties and the step difficulties, by allowing for residual variation.

To develop polytomous item explanatory models with random item errors based on the PCM, it is helpful to clarify the target polytomous item difficulty parameters which are explained by item properties as well as the three types of polytomous random item errors which are added to the models. First, using the MFRM approach, the item location parameters β_i can be explained by item properties and then we can add (a) overall item random error for residual variation in the overall item difficulties. Next, using the LPCM approach, the step difficulty parameters δ_{im} can be explained by item properties and then we can add (b) step specific random item error or (c) item-step random error for residual variation in the step difficulties. Table 4-2 presents an overarching framework summarizing the polytomous item explanatory models with random item errors that I propose with model specifications in this paper.

Many-Facet Rasch Model with Overall Item Random Error (MFRM + OIE)

The first model is referred to as the Many-Facet Rasch Model with Overall Item Random Error (MFRM + OIE). By using the twofold item parameterization for the PCM, an overall item random error term ϵ_i is added to the MFRM-based polytomous item explanatory model in Equation 4.9, formulated as:

$$ln\frac{Pr(y_{pi}=m|\theta_p,\epsilon_i)}{Pr(y_{pi}=m-1|\theta_p,\epsilon_i)}=\theta_p-\sum_{k=0}^{K}\gamma_k x_{ik}-\epsilon_i-\tau_{im}, \quad (4.12)$$

where $\epsilon_i \sim N(0,\sigma_\epsilon^2)$, $\theta_p \sim N(0,\sigma_\theta^2)$, $\tau_{i0}=0$, and $\sum_{m=1}^{M_i}\tau_{im}=0$. Also, where ϵ_i is an error or residual term for the overall item difficulties with a normal distribution of a zero mean and a variance of σ_ϵ^2, γ_0 is the item intercept representing the difficulty for items with all x_{ik} = 0 for k > 0, γ_k is the regression weight or the effect of item property k on the overall item difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, x_{ik} is the value of item i on item property k, and τ_{im} is a step deviation parameter for the m-th step of item i. For model identification, the mean person ability is constrained to zero. Note that this model with no item property predictors ($K=0$) is formally equivalent to the overall item random effects model in Equation 4.5, that is, $\gamma_0=\mu_\beta$.

Table 4-2. Three proposed polytomous item explanatory models with random item errors

Item Parameteri-zation	Target Item Difficulty Parameter	Relevant Explanatory Approach	Polytomous Random Item Error	Model Specification
Two fold $\delta_{im}=\beta_i+\tau_{im}$	Item Location (Overall Item Difficulty) Parameter β_i	Many-Facet Rasch Model $\beta_i'=\sum_{k=0}^{K}\gamma_k x_{ik}$	Overall Item Random Error ϵ_i	$\delta_{im}=\sum_{k=0}^{K}\gamma_k x_{ik}+\epsilon_i+\tau_{im}$ $\epsilon_i \sim N(0,\sigma_\epsilon^2)$ $\sum_{m=1}^{M_i}\tau_{im}=0$
One fold δ_{im}	Step Difficulty Parameter δ_{im}	Linear Partial Credit Model $\delta_{im}'=\sum_{k=0}^{K}\omega_{km}x_{ik}$	Step Specific Random Item Error ξ_{im}	$\delta_{im}=\sum_{k=0}^{K}\omega_{km}x_{ik}+\xi_{im}$ $\boldsymbol{\xi_i}\,(\xi_{i1},\ldots,\xi_{im})' \sim MVN_m(\mathbf{0},\boldsymbol{\Sigma})$
			Item-Step Random Error ε_{im}	$\delta_{im}=\sum_{k=0}^{K}\omega_{km}x_{ik}+\varepsilon_{im}$ $\varepsilon_{im} \sim N(0,\sigma_\varepsilon^2)$

Note. x_{ik} is the value of item i on item property k, γ_k is the regression weight or the effect of item property k on the overall item difficulties, $\tau_{im}$is a step deviation parameter for the m-th step of item i, and ω_{km} is the regression weight or the effect of item property k on the m-th step difficulties, ϵ_i is an error term for residual variation in the overall item difficulties, $\boldsymbol{\xi_i}\,(\xi_{i1},\xi_{i2},\ldots,\xi_{im})'$ is an error term vector for step specific residual variation in the step difficulties, and ε_{im} is an error term for item-step residual variation in the step difficulties.

Linear Partial Credit Model with Step Specific Random Item Error (LPCM + SSE)

The second model is referred to as the Linear Partial Credit Model with Step Specific Random Item Error (LPCM + SSE). By using the original PCM's item parameterization, a vector of the step specific random item error term $\boldsymbol{\xi_i}$ $(\xi_{i1}, \xi_{i2}, \dots, \xi_{im})$' is added to the LPCM-based polytomous item explanatory model in Equation 4.11, formulated as:

$$ln\frac{Pr(y_{pi} = m|\theta_p, \boldsymbol{\xi_i})}{Pr(y_{pi} = m-1|\theta_p, \boldsymbol{\xi_i})} = \theta_p - \sum_{k=0}^{K} \omega_{km} x_{ik} - \xi_{im}\,, \tag{4.13}$$

where

$$\boldsymbol{\xi_i} = \begin{bmatrix} \xi_{i1} \\ \xi_{i2} \\ \vdots \\ \xi_{im} \end{bmatrix} \sim MVN_m(\mathbf{0}, \boldsymbol{\Sigma}),\ \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{\xi 1} & \sigma_{\xi 1 \xi 2} & \cdots & \sigma_{\xi 1 \xi m} \\ \sigma_{\xi 2 \xi 1} & \sigma^2_{\xi 2} & \cdots & \sigma_{\xi 2 \xi m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\xi m \xi 1} & \sigma_{\xi m \xi 2} & \cdots & \sigma^2_{\xi m} \end{bmatrix},\quad \theta_p \sim N(0, \sigma^2_\theta),\ \text{and}$$

$\omega_{k0} = 0$. Also, where $\boldsymbol{\xi_i}$ is an error or residual term vector for the step difficulties following a multivariate normal (MVN) distribution with a zero mean vector $\mathbf{0}$ and a variance-covariance matrix $\boldsymbol{\Sigma}$ $(m \times m)$, ω_{0m} is the step intercept representing the m-th step difficulty for items with all x_{ik} = 0 for k > 0, ω_{km} is the regression weight or the effect of item property k on the m-th step difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, and x_{ik} is the value of item i on item property k. For model identification, the mean person ability is constrained to zero. Note that this model with no item property predictors ($K = 0$) is formally equivalent to the step specific random item effects model in Equation 4.6, that is, $\omega_{0m} = \mu_{\delta m}$ for every m-th step.

Linear Partial Credit Model with Item-Step Random Error (LPCM + ISE)

The third model is referred to as the Linear Partial Credit Model with Item-Step Random Error (LPCM + ISE). By using the original PCM's item parameterization, an item-step random error term ε_{im} is added to the LPCM-based polytomous item explanatory model in Equation 4.11, formulated as:

$$ln\frac{Pr(y_{pi} = m|\theta_p, \varepsilon_{im})}{Pr(y_{pi} = m-1|\theta_p, \varepsilon_{im})} = \theta_p - \sum_{k=0}^{K} \omega_{km} x_{ik} - \varepsilon_{im}, \tag{4.14}$$

where $\varepsilon_{im} \sim N(0, \sigma^2_\varepsilon)$, $\theta_p \sim N(0, \sigma^2_\theta)$, and $\omega_{k0} = 0$. Also, where ε_{im} is an error or residual term for the step difficulties with a normal distribution of a zero mean and a variance of σ^2_ε, and other parameters are defined as before in the second model. For model identification, the mean person ability is constrained to zero. Note that this model with no item property predictors ($K = 0$) is not equivalent to the item-step random effects model in Equation 4.7, that is, $\omega_{0m} \neq \mu_\delta$. To be

identical, values of the step intercepts should be collapsed across steps into the one same value ($\omega_{01} = \cdots = \omega_{0m} = \omega_0 = \mu_\delta$). This is because the item-step random effects do not adequately reflect the multi-categorical scale structure in the ordinal items. Nonetheless, the item-step random effects are used to allow for residual variation in the step difficulties.

The step specific random item effects are more general than the item-step random effects which have additional assumptions of the same mean and variance for every step and the zero correlation between steps. In this sense, the LPCM + SSE is a more general LPCM-based polytomous item explanatory model than the LPCM + ISE. Therefore, for simulation studies, I will use the LPCM + SSE as a data generating model to simulate data for the LPCM-based polytomous item explanatory models with and without random item errors in different explanatory conditions.

4.4. Estimation of Polytomous Random Item Effects Models

4.4.1. Estimation of Crossed Random Effects Models

The three proposed polytomous item explanatory models with random item errors essentially have crossed random effects. Given the fact that item responses are nested in persons and also in items but these two sides of the response data are not nested (Van den Noortgate, De Boeck, & Meulders, 2003), allowing for random effects on both item and person sides makes them crossed, as illustrated in Figure 4-1. However, estimation of crossed random effects models in IRT is demanding due to the complexity and difficulty in numerical integration (Cho & Rabe-Hesketh; 2011; De Boeck & Wilson, 2004). This becomes exacerbated and practically difficult for polytomous random item effects models using adjacent-categories logits.

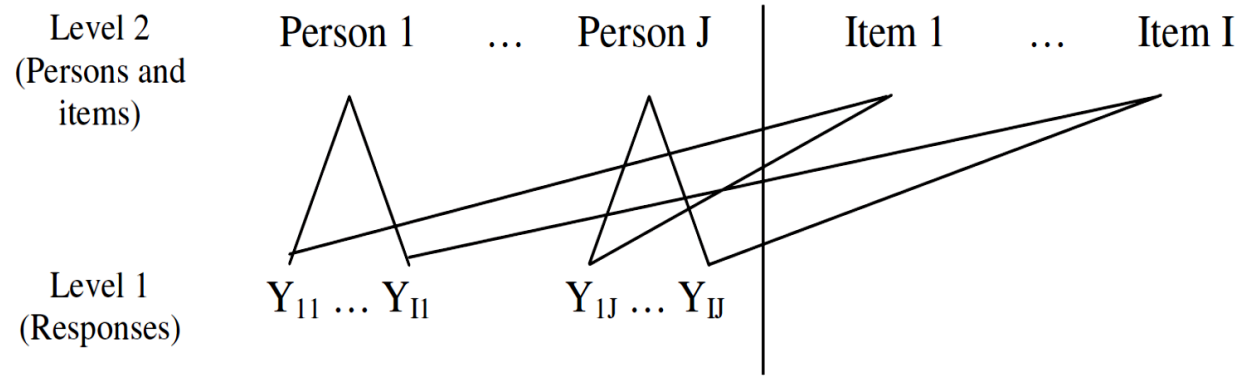


Figure 4-1. Cross-classified item responses in a random item effects model. Adapted from "Cross-classification multilevel logistic models in psychometrics," by W. Van den Noortgate, P. De Boeck, and M. Meulders, 2003, *Journal of Educational and Behavioral Statistics*, 28(4), p. 372. Copyright 2003 by American Educational Research Association.

A common estimation approach for IRT models is maximum likelihood estimation (MLE). For statistical inference, model parameters are estimated by maximizing the likelihood function of the marginalized model, which is marginal maximum likelihood estimation (MML; Bock and

Aitkin, 1981; Tuerlinckx, Rijmen, Verbeke, & De Boeck, 2006). The likelihood is marginal over the random effects. However, the likelihood function cannot be written in a closed form and the integrals are intractable (Tuerlinckx et al., 2004), so numerical integration (e.g. adaptive quadrature) is typically used. The following three approaches are most commonly used to handle the intractable integrals in IRT models, including polytomous item response models (Tuerlinckx et al., 2006): (a) numerical approximation to the integral such as (adaptive) Gaussian quadrature, Monte Carlo integration, and (b) approximation of the integrand such as the Laplace approximation method and quasi-likelihood approaches. Even so, for crossed random effects, burdensome integrals in the marginal likelihood become high-dimensional and numerical integration becomes prohibitive (Cho, De Boeck, Embretson, & Rabe-Hesketh, 2014; Janssen, Schepers, & Peres, 2004).

To overcome such intractable and demanding integration for random item IRT models with crossed random effects, there are three typical approaches to model parameter estimation. (Cho, Partchev, & De Boeck, 2012; Cho et al., 2014): (a) Bayesian estimation methods using Markov chain Monte Carlo (MCMC) algorithm (Fox, 2010), (b) approximation methods such as the Laplace approximation (Bellio & Varin, 2005), and (c) an approach with maximization divided into separate steps with possible alternation between them, such as the alternating imputation posterior (AIP) method (Cho & Rabe-Hesketh, 2011).

The first approach is a well-known Bayesian method, Monte Carlo simulation-based parameter estimation using MCMC algorithm. Briefly, the MCMC algorithm approximates the posterior distribution of a model using successive samples of parameter values that are randomly drawn from the theoretical distribution. This is done by means of a sampling algorithm such as a Gibbs sampler which creates a Markov chain through successively sampling from a set of full conditional distributions of individual parameters over a large number of iterations (Fox, 2010; Janssen, Schepers, & Peres, 2004). The Markov chain approximates the joint posterior distribution of the parameters of the model, from which the parameters are estimated as the posterior means or medians. This simulation-based estimation is flexible and useful to estimate complicated statistical models tailored to specific research questions. (Johnson & Sinharay, 2005; Tuerlinckx et al., 2006), and it is practical to estimate all model parameters from the entire posterior distributions in one step (Tuerlinckx et al., 2004). Bayesian methods which consider all effects as random parameters are a straightforward approach to estimation of the crossed random effects (Cho et al., 2014). However, Bayesian methods have several drawbacks such as expensive computation (for crossed random effects, numerical integration is far more expensive), the difficulty of specifying vague priors for parameters and convergence assessment, and unavailability in general statistical analysis software (Bellio & Varin, 2005; Cho et al., 2014; Tuerlinckx et al., 2006). Some examples of available software for Bayesian methods are *Stan* (Carpenter et al., 2017), MLwiN (Rasbash et al., 2000), WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2010), JAGS (Plummer, 2003), *MCMCpack* R package (Martin, Quinn, & Park, 2011), PROC MCMC in SAS (SAS Institute Inc., 2013), and *Bayesmh* command in Stata (StataCorp., 2015).

The second approach is approximation of the integrands or models that make the integrals faster to compute. Important examples are the Laplace approximation and quasi-likelihood methods (Breslow & Clayton, 1993; Cho, Partchev, & De Boeck, 2012; Goldstein & Rasbash, 1996; Van den Noortgate, De Boeck, & Meulders, 2003). It is known that use of the quasi-likelihood extensions such as marginal quasi-likelihood (MQL) and penalized quasi-likelihood

(PQL) performs poorly with a downward bias in the estimated variance components (Cho & Rabe-Hesketh, 2011). The Laplace approximation appears to perform better (Bellio & Varin, 2005) but also performs poorly in some situations (Joe, 2008). These approximation methods are readily implemented using general statistical analysis software, however, they often do not provide consistent estimation, particularly for binary data in small clusters (Bellio & Varin, 2005; Joe, 2008). Some examples of available software for this approach are SAS macro GLIMMIX (SAS Institute Inc., 2013), HLM (Raudenbush, Bryk, & Congdon, 2004), MLwiN (Rasbash et al., 2000), *lme4* R package (Bates, Maechler, Bolker, & Walker, 2014), and *xtmelogit / meglm* commands in Stata (StataCorp., 2015).

A typical example of the third approach for estimation of crossed random effects models is the alternating imputation posterior (AIP) method, which implements maximization divided into separate steps or models with possible iteration and alternation between them (Cho & Rabe-Hesketh, 2011). The AIP method uses the alternating imputation posterior algorithm which iterates data augmentation of imputing missing data in the I-step and updating the approximation of the posterior distribution in the P-step. To obtain approximated maximum likelihood estimates, this algorithm alternates between the divided sub-models (wings) until convergence, by iterative sampling from the respective conditional posterior distributions estimated using an approximate MLE method for numerical approximation to the integral (Cho, Partchev, & De Boeck, 2012; Cho & Rabe-Hesketh, 2011; Clayton & Rasbash, 1999). Compared to Bayesian methods, the AIP method does not require priors for model parameters and converges rapidly due to simultaneously updating the parameters. However, it requires iterative sampling and relies on asymptotic theory in sample size, and it is difficult to assess convergence (Cho & Rabe-Hesketh, 2011). The AIP method is implemented using general statistical analysis software but requires additional complicated programming.

Estimation of polytomous random item effects models with crossed random effects is more difficult in practice. The three proposed polytomous item explanatory models with random item errors have been developed by adapting the PCM, which uses an adjacent-categories logit link rather than logistic link in the statistical model. However, the adjacent-categories logit link is not common in general statistical analysis software for maximum likelihood-based estimation methods. In particular, it is rare that software can estimate models with crossed random effects for polytomous data. For example, the *lmer* function in the *lme4* R package (De Boeck et al., 2011) and the *xtmelogit / meglm* commands in Stata (StataCorp, 2015), which use the Laplace approximation, can fit logistic crossed random effects models to dichotomous data but cannot fit adjacent-categories logit crossed random effects models to polytomous data. Moreover, most of the IRT analysis software such as PARSCALE (Muraki & Bock, 1997) and BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) cannot estimate crossed random effects IRT models.

4.4.2. Bayesian methods for Polytomous Item Explanatory Models with Random Item Errors

To estimate the proposed polytomous item explanatory models with random item errors for simulated and real data, I will use a Bayesian inference via MCMC algorithm. It is a straightforward approach to estimation of the crossed random effects (Cho et al., 2014), and also flexible and useful to estimate complicated models (Johnson & Sinharay, 2005; Cho, Partchev, &

De Boeck, 2012). Although it is not MLE, it can yield very similar results as other comparable methods such as the AIP and the Laplace approximation (Cho, Partchev, & De Boeck, 2012; Cho & Rabe-Hesketh, 2011; De Boeck, 2008; Janssen, Schepers, & Peres, 2004). In fact, Bayesian methods are getting to be more commonly used (e.g. Ames & Samonte, 2015; Briggs & Wilson, 2007; Curtis, 2010; Fox, 2010; Kang, Cohen, & Sung, 2009; Lathrop & Cheng, 2017; Luo & Jiao, 2018).

To implement the MCMC, the *RStan* R package (Stan Development Team, 2018), the R implementation of Stan (Carpenter et al., 2017), is used in this research. Stan is a newly developed Bayesian program that implements a No-U-Turn sampler (NUTS) based on a Hamiltonian Monte Carlo (HMC) sampling algorithm. NUTS is more efficient and robust than Gibbs sampling or the Metropolis-Hastings algorithm to explore the posterior parameter space (Carpenter et al., 2017; Luo & Jiao, 2018). As model size and complexity grow, Stan is faster and more robust than previous Bayesian software such as the WinBUGS (Lunn et al., 2000) for models with complex posteriors[12] (Luo & Jiao, 2018; Monnahan, Thorson, & Branch, 2017).

Stan provides full Bayesian inference for model parameters treated as random variables. Bayesian inference requires a set of prior distributions of the model parameters that could reflect our beliefs or a priori knowledge about those parameters before some evidence is taken into account (Fox, 2010). The priors may be informative, weakly informative, or non-informative (Luo & Jiao, 2018). To estimate the three proposed polytomous random item effects models using Stan, I will use the following priors and hyperpriors for the model parameters in each model.

Priors of Model Parameters for the MFRM + OIE

$\theta_p \sim N(0, \sigma_\theta^2)$,
$\sigma_\theta \sim Cauchy(0, 5)$, truncated at 0,
$\gamma_k \sim N(0, 10^2)$,
$\tau_{im} \sim N(0, 3^2)$,
$\epsilon_i \sim N(0, \sigma_\epsilon^2)$,
$\sigma_\epsilon \sim Cauchy(0, 5)$, truncated at 0.

Priors of Model Parameters for the LPCM + SSE

$\theta_p \sim N(0, \sigma_\theta^2)$,
$\sigma_\theta \sim Cauchy(0, 5)$, truncated at 0,
$\omega_{km} \sim N(0, 10^2)$ for $m \in 1{:}M$,
$\boldsymbol{\xi_i} \sim MVN_m(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\xi_i} = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{im})'$,
$\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma_\xi}) \times \boldsymbol{\Omega} \times \text{diag}(\boldsymbol{\sigma_\xi})$, where $\boldsymbol{\sigma_\xi} = (\sigma_{\xi 1}, \sigma_{\xi 2}, \ldots, \sigma_{\xi m})'$,
$\sigma_{\xi m} \sim Cauchy(0, 2.5)$ for $m \in 1{:}M$, truncated at 0,
$\boldsymbol{\Omega} \sim LKJ(1)$.

[12] In the pilot simulation study I conducted, as in the findings by Monnahan, Thorson, and Branch (2017), Stan outperformed WinBUGS to estimate the LPCM + SSE which is the most complex polytomous random item effects model. It appeared that fitting the model to simulated data could not converge using WinBUGS with 4 chains and 7,000 iterations but it could converge using Stan with 4 chains and 1,000 iterations.

Priors of Model Parameters for the LPCM + ISE

$$\theta_p \sim N(0, \sigma_\theta^2),$$
$$\sigma_\theta \sim Cauchy(0, 5), \text{ truncated at 0,}$$
$$\omega_{km} \sim N(0, 10^2) \text{ for } m \in 1{:}M,$$
$$\varepsilon_{im} \sim N(0, \sigma_\varepsilon^2),$$
$$\sigma_\varepsilon \sim Cauchy(0, 5), \text{ truncated at 0.}$$

The prior for the person ability parameter θ_p in all models is a normal distribution with zero mean. The hyperprior for the standard is a Cauchy distribution truncated to the positive real values, i.e., a half-Cauchy distribution, with a location of zero and a scale of 5. This half-Cauchy is a weakly informative prior that provides a large mass near the boundary and also flattens out at the boundary, regarded as a good default choice for scale parameters (Lathrop & Cheng, 2017; Stan Development Team, 2017). In addition, the priors for the fixed item effects parameters, the item property effect parameters γ_k and ω_{km} and step deviation parameter τ_{im}, have weakly informative priors. Each parameter follows a normal distribution with a zero mean (for parameter identifiability) and a fairly large standard deviation (for flatter priors). For the random item effects parameters ϵ_i, ξ_{im}, and ε_{im}, each of the priors is a normal distribution with a zero mean as a usual error, and a half-Cauchy distribution is specified as a weakly informative hyperprior for the standard deviation of each prior.

In particular, the step specific random item error vector $\boldsymbol{\xi_i}$ $(\xi_{i1}, \xi_{i2}, \ldots, \xi_{im})$' in the LPCM + SSE follows a multivariate normal (MVN) distribution with a zero mean vector $\mathbf{0}$, and the prior for the variance-covariance matrix $\boldsymbol{\Sigma}$ $(m \times m)$ is decomposed into a scale or standard deviation vector $\boldsymbol{\sigma_\xi}$ $\left(\sigma_{\xi 1}, \sigma_{\xi 2}, \ldots, \sigma_{\xi m}\right)'$ and a correlation matrix $\boldsymbol{\Omega}$. The components of the scale vector $\sigma_{\xi m}$ have a weakly informative hyperprior in each step m that follows a half-Cauchy distribution with a small scale of 2.5, as recommend by the Stan manual (Stan Development Team, 2017). It is also recommended that an LKJ (Lewandowski, Kurowicka, & Joe, 2009) prior is placed on the correlation matrix $\boldsymbol{\Omega}$, and the LKJ correlation distribution with a shape of one is weakly informative due to reducing to the identity distribution. Since it is computationally more efficient and arithmetically stable to use a Cholesky factor of the variance-covariance matrix $\boldsymbol{\Sigma}$ (Stan Development Team, 2017), the Cholesky decomposition of $\boldsymbol{\Sigma}$ is also implemented. Lastly, although the fixed and random item effects are step specific in the LPCM + SSE, a univariate normal distribution is used as the prior for the item property effect parameter ω_{km} because it produces the same density much more efficiently to the weakly informative multivariate normal prior with no correlation information in Stan (Stan Development Team, 2017).

These priors are justified by the Stan manual and other Bayesian IRT studies (e.g. Curtis, 2010; Fox, 2010; Kang, Cohen, & Sung, 2009; Lathrop & Cheng, 2017; Luo & Jiao, 2018). Stan syntax for the three polytomous item explanatory models with random item errors, developed by adapting the simple PCM syntax by Furr (2017a), is shown in the Appendix A.3.

4.5. Simulation Studies

4.5.1. Simulation Design and Data Generation

The goal of this research is to extend the LLTM + ε approach to polytomous items and to demonstrate practical applications of the extended models for empirical data sets. I first addressed the two polytomous item explanatory models using the MFRM and the LPCM approaches and then investigated the three polytomous item explanatory models with random item errors: the MFRM + OIE, the LPCM + SSE, and the LPCM + ISE. Before applying the newly proposed models to real data, it would be helpful to examine how those item explanatory models work in different explanatory conditions using simulation methods.

To assess model comparison and parameter recovery as well as effects of model misspecification on parameter estimation, two simulation studies were conducted. The first simulation, referred to as a model comparison study, aimed to compare the performance of the three polytomous random item effects models with the two fixed item property effects models as well as with the PCM in each polytomous item explanatory approach. To investigate the relative performance in an ordinary measurement context that the empirical data sets are observed once, the models were fitted to the data simulated with one replication. The relative performance between the fitted models was evaluated by comparing goodness-of-fit and by examining agreement for the estimated and calculated step difficulties. The second simulation investigated model parameter recovery when the simulated datasets were analyzed using a data generating model in each polytomous item explanatory approach. To check parameter recovery, the simulation was replicated many times so that bias and root mean squared error could be estimated. In both simulation studies, the effects of model misspecification on parameter estimation were also examined when the simulated datasets were analyzed using the other models rather than the generating model in each polytomous item explanatory approach.

In particular, to set different explanatory conditions in simulating polytomous item explanatory datasets, I considered two conditions of the extent to which the incorporated item properties explain or predict the polytomous item difficulties: a "low information" condition and a "high information" condition (Furr, 2017b). Because an item explanatory model has an item regression component, the concept of the coefficient of determination, R^2, can be employed. R^2 is defined as the proportion of the total item variance that is explained by the item predictors:

$$R^2 = \frac{v_{model}^2}{v_{model}^2 + r_{error}^2}, \tag{4.15}$$

where v_{model}^2 is the variance of the weighted sums of the item property effects and r_{error}^2 is the item error (or residual) variance. I set $R^2 = 0.3$for the low information condition and $R^2 = 0.9$ for the high information condition. This range of R^2 value contains empirical findings using the LLTM + ε (e.g. 0.396 from Harting and colleagues' study in 2012; 0.875 from De Boeck's study in 2008). These conditions reflect the degree of uncertainty in explanation and/or prediction of the random item effects. Since there is no "perfect" explanation by observed item predictors in practice, R^2 cannot be 1.00.

In order to make the simulation design more substantive, I supposed a reading comprehension test in the context of item explanatory measurement situations. Grounded on empirical evidence from the reading comprehension literature, two categorical item properties and one continuous item property were created: (a) two levels of reading comprehension—local and global comprehension (Hartig et al., 2012), (b) three types of reasoning in narrative comprehension—literal, inferential, and evaluative reasoning (Basaraba, Yovanoff, Alonzo, & Tindal, 2013), and (c) text complexity in terms of vocabulary and grammatical structures (Hartig et al., 2012), which was assumed to use a 5-point interval scale (very simple = 0, simple = 1, moderate = 2, complex = 3, very complex = 4). By designing items based on the three item properties, a total of 30 items were generated to fit a 2 × 3 × 5 design with one item within each cell. It was assumed that a total of 500 persons responded to the 30 items; their answers were scored to have three ordered-category responses and hence two steps of adjacent categories ($m = 1, 2$). When incorporating these item properties into the polytomous item explanatory models, the two categorical item properties were dummy coded so that the local comprehension and the literal reasoning are a reference for each item property.

Note that the MFRM and the LPCM are conceptually and functionally different item explanatory approaches. The target difficulty parameters of polytomous items which are explained by item properties, the types of item property effects, and the types of polytomous random item errors are different between the two. The overall item difficulties are determined by the item property effects and overall item random error in the MFRM approach. Yet, the step difficulties are determined by the step specific item property effects and either step specific random item error or item-step random error in the LPCM approach. In light of the differences between the two item explanatory approaches to polytomous data, two different data generating models (the MFRM + OIE and the LPCM + SSE) were used in both simulation studies. This is because they are the most general models that include random item effects in each polytomous item explanatory approach.

Data Generation using the MFRM + OIE

In the MFRM approach to explanation or prediction of the overall item difficulties, using the MFRM + OIE, two sets of polytomous item response data were generated for the two information conditions ($R^2 = 0.3$ and 0.9). Grounded on empirical evidence for each item property (see Basaraba et al, 2013; Hartig et al., 2012), generating values were chosen for the item property effect parameters. Due to the model constraint of a zero sum of step deviations for each item in Equation 4.12 ($\sum_{m=1}^{2} \tau_{im} = 0$), generating values of the step deviation parameters were made only for the first step. These values for the fixed item effects parameters were the same across the two conditions. For the overall item error variance, using Equation 4.15, a generating value was calculated for each information condition. Since the Stan software uses a scale for distribution commands, a standard deviation value was inputted and provided. To sample the latent abilities for 500 persons from a normal distribution, a standard deviation of 1.5 was used in all conditions. Table 4-3 presents true generating values for the parameters in the MFRM + OIE.

Table 4-3. Parameter generating values for the MFRM + OIE

Information condition		$R^2 = 0.3$ (Low)	$R^2 = 0.9$ (High)
Model parameter		Generating value	
Fixed effects			
Intercept	γ_0	− 1.00	− 1.00
Global comprehension	γ_1	0.80	0.80
Evaluative reasoning	γ_2	0.60	0.60
Inferential reasoning	γ_3	0.40	0.40
Text complexity	γ_4	0.20	0.20
Step deviation (1st step)	τ_{i1}	− 0.10 for $i \in 1,3,\ldots,29$, − 0.20 for $i \in 2,4,\ldots,30$	− 0.10 for $i \in 1,3,\ldots,29$, − 0.20 for $i \in 2,4,\ldots,30$
Step deviation (2nd step)	τ_{i2}	$\tau_{i2} = -\tau_{i1}$	$\tau_{i2} = -\tau_{i1}$
Random effects			
Overall item error variance	σ_ϵ^2	0.85^2	0.19^2
Person variance	σ_θ^2	1.50^2	1.50^2

Data Generation using the LPCM + SSE

In the LPCM approach to explanation or prediction of the step difficulties, two sets of polytomous item response data were generated using the LPCM + SSE with the two information conditions ($R^2 = 0.3$ and 0.9). Since the fixed item effects parameters are the step specific item property effects in this approach, generating values were made for each step across the item properties based on the empirical evidences (e.g. Basaraba et al, 2013; Hartig et al., 2012). These values were the same between the two conditions. For the step specific random item error variances, a generating value for each step was calculated using Equation 4.15 in each information condition. A correlation of the item errors between the steps was set to zero because this correlation is nothing to do with R^2 for each step and I considered the weakly informative multivariate normal prior with no correlation information for the item property effects in the LPCM + SSE. For the person random effects, simulation conditions were the same with the MFRM approach. Table 4-4 shows true generating values for the parameters in the LPCM + SSE.

Table 4-4. Parameter generating values for the LPCM + SSE

Information condition		$R^2 = 0.3$ (Low)	$R^2 = 0.9$ (High)
Model parameter		Generating value	
Fixed effects			
Intercept	ω_{01}	− 1.50	− 1.50
	ω_{02}	− 0.50	− 0.50
Global comprehension	ω_{11}	0.50	0.50
	ω_{12}	1.10	1.10
Evaluative reasoning	ω_{21}	0.40	0.40
	ω_{22}	0.80	0.80
Inferential reasoning	ω_{31}	0.20	0.20
	ω_{32}	0.60	0.60
Text complexity	ω_{41}	0.10	0.10
	ω_{42}	0.30	0.30
Random effects			
Item error variance (1st step)	$\sigma_{\xi 1}^2$	0.51^2	0.11^2
Item error variance (2nd step)	$\sigma_{\xi 2}^2$	1.20^2	0.26^2
Correlation between the steps	$\rho_{\xi 12}$	0.00	0.00
Person variance	σ_θ^2	1.50^2	1.50^2

4.5.2. Study 1: Model Comparison

Analysis

A model comparison study was conducted to examine how the proposed polytomous item explanatory models with random item errors (the MFRM + OIE, the LPCM + SSE, and the LPCM + ISE) perform in comparison with the fixed item property effects models (the MFRM and the LPCM) and the PCM in the low information ($R^2 = 0.3$) and the high information ($R^2 = 0.9$) conditions. Since the two item explanatory approaches to polytomous items are methodologically different, a model comparison simulation was conducted by using each approach. The six models were fitted to the two datasets generated for the two information conditions using a data generating model in each item explanatory approach-based simulation.

To evaluate the relative performance of the proposed models, I compared goodness-of-fit of the fitted models. Since I used full Bayesian methods for parameter estimation using Stan, three Bayesian goodness-of-fit indices were used: the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002, 2014), the widely applicable information criterion (WAIC; Watanabe, 2010), and an information criterion version of the leave-one-out cross-validation (LOOIC; Vehtari, Gelman, & Gabry, 2017; Vehtari, Gelman, Gabry, & Yao,

2018). These Bayesian measures of model fit and adequacy are widely used to assess the generalization utility or the predictive ability of the candidate models (Piironen & Vehtari, 2017). In particular, WAIC and LOOIC are theoretically justified and empirically superior to traditional model selection criteria such as the Akaike information criterion and the Bayesian information criterion (Luo & Jiao, 2018; Piironen & Vehtari, 2017; Spiegelhalter et al., 2014). I computed DIC using the log-likelihood from the Stan output, and the *loo* R package (Vehtari et al., 2018) was used to compute WAIC and LOOIC. It should be noted that all these methods are conditional on the person abilities and item random effects and it might be preferable to use methods based on marginal likelihoods (Furr, 2017b). I also examined agreement for the estimated and calculated step difficulties between the fitted models using correlation coefficients and graphical comparisons (see Fischer, 1973; Holling, Bertling, & Zeuch, 2009). The directly estimated step difficulties in the PCM were compared to the calculated step difficulties in the two polytomous item explanatory approaches.

In addition to the model comparison, I investigated how model misspecification affects parameter estimation, when using a model other than the generating model in each item explanatory approach. Focusing on the item property effect parameters, the true generating values were compared to the parameter estimates with 95% credible intervals (Bayesian confidence intervals).

The R software (R Development Core Team, 2013) was used to generate a series of datasets, and the *RStan* R package (Stan Development Team, 2018) was used to analyze the generated datasets. Four chains of 1,000 iterations were used for MCMC estimation. A total of 2,000 draws across the four chains were used to estimate posterior means of the parameters after discarding the first 500 burn-in (or warm-up) iterations of each chain. Convergence of the four chains was monitored by checking the estimated potential scale reduction statistic $\hat{R}$ (Gelman & Rubin, 1992). A value of the $\hat{R}$ statistic near 1 usually indicates model convergence. The value of 1.1 has been recommended as a threshold and $\hat{R} < 1.1$ is desirable (Gelman & Rubin, 1992; Luo & Jiao, 2018; Stan Development Team, 2017). For additional convergence diagnostics, I also checked the effective sample size for a chain, which is an estimate of the number of independent draws that would lead to the same expected precision (see Carpenter et al., 2017; Stan Development Team, 2017).

Results

Many-Facet Rasch Model approach-based simulation

In the MFRM approach-based model comparison simulation, the five polytomous item explanatory models with and without random item errors and the PCM were fitted to two sets of the MFRM + OIE generation data with the two information conditions. Regarding convergence, the $\hat{R}$ statistic values were less than 1.03 and the effective sample sizes were large enough in all fitted models and information conditions.

Table 4-5 presents the goodness-of-fit results of the six models and Figure 4-2 depicts them graphically. It appeared that the order of DIC, LOOIC, and WAIC across the models was consistent within each information condition. Although the data were generated using the MFRM + OIE, the LPCM + SSE appeared to be the best fitting model in both information conditions. This makes sense because the LPCM + SSE has the most complex and general random item effects. The LPCM

Table 4-5. Goodness-of-fit of models fitted to the MFRM + OIE generation data

R^2	Model	k	DIC	LOOIC	WAIC
0.3	PCM	61	21308.78	21324.55	21317.04
	MFRM	36	23872.70	23881.62	23875.24
	MFRM + OIE*	37	21312.05	21325.29	21317.65
	LPCM	11	23912.29	23919.63	23913.18
	LPCM + SSE	14	21301.17	21312.64	21305.36
	LPCM + ISE	12	21315.83	21326.82	21319.57
0.9	PCM	61	22893.63	22907.70	22900.09
	MFRM	36	23225.41	23234.96	23228.02
	MFRM + OIE*	37	22898.96	22910.96	22903.28
	LPCM	11	23212.11	23221.20	23214.27
	LPCM + SSE	14	22877.48	22889.05	22881.30
	LPCM + ISE	12	22886.04	22896.33	22888.92

Note: * Data generating model; k = The number of estimated parameters.

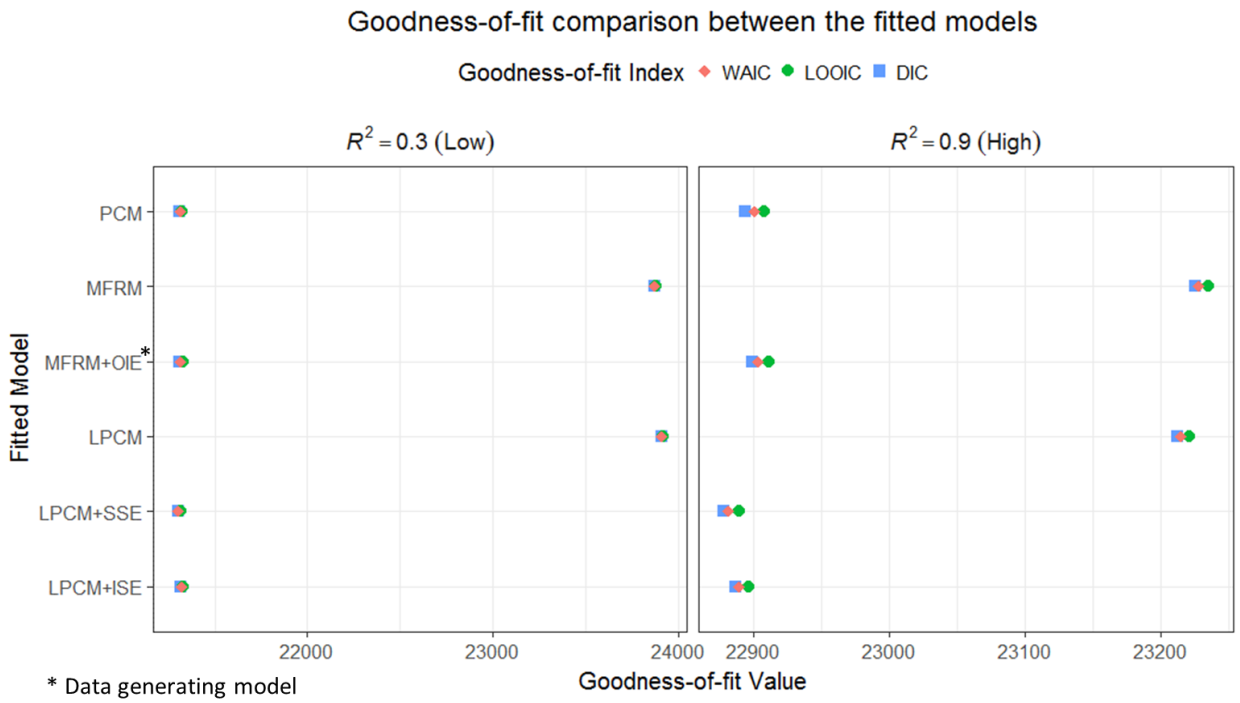


Figure 4-2. Goodness-of-fit of models fitted to the MFRM + OIE generation data

+ SSE even fit better than the PCM regarded as the saturated model in the fixed item effects models. This result corresponds to other applications of the LLTM + ε. De Boeck (2008) reported the smaller value of the Bayesian information criterion for the homoscedastic LLTM + ε than the random person fixed item (RPFI) Rasch model. The LPCM + ISE appeared to fit worse than the PCM in the low information condition but fit better in the high information condition. Despite the

data generating model, the MFRM + OIE appeared to fit worse than the PCM in both information conditions.

The goodness-of-fit performance of the MFRM and the LPCM was inferior to that of the PCM in both information conditions. This is an expected finding because item explanatory models with fixed item effects usually fit worse than the saturated model (e.g. Kubinger, 2009; Tuerlinckx & Wang, 2004). A methodological advantage of a smaller number of item parameters in the LLTM approach is at the cost of statistically lower goodness-of-fit.

Compared to the two fixed item property effects models, the three random item effects models showed a significantly superior goodness-of-fit in both information conditions. This indicates that random item errors account for residual variation in the polytomous item difficulties as in the LLTM + ε applications (see De Boeck, 2008; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009). The LLTM + ε approach compensates for the loss of goodness-of-fit due to linear restrictions on the item parameters while keeping the methodological advantage of an item explanatory model.

A correlation analysis was conducted to examine the practical significance (effect size) of the models. Correlations were estimated between the directly estimated step difficulties in the PCM and the calculated step difficulties in the five polytomous item explanatory models (see Table 4-6). In both information conditions, the three random item effects models provided almost perfect agreement with the PCM in that correlations were close to 1. This implies that item error terms worked well to enhance prediction of the overall item difficulties or the step difficulties by allowing for residual variation in both information conditions. However, the two fixed item property effects models provided similar agreement within each information condition but lower agreement in the low information condition (0.66 for the MFRM and 0.68 for the LPCM) and higher agreement in the high information condition (0.91 for the MFRM and 0.90 for the LPCM). The lower R^2 suggests more unexplained residual variability, which cannot be predicted by the fixed item property effects without item error terms.

Table 4-6. Correlations of the estimated and calculated step difficulties between the models fitted to the MFRM + OIE generation data

R^2	Model	PCM	MFRM	MFRM + OIE	LPCM	LPCM + SSE	LPCM + ISE
0.3	PCM	1	0.66	1	0.68	1	1
	MFRM		1	0.67	0.96	0.66	0.67
	MFRM + OIE*			1	0.68	1	1
	LPCM				1	0.69	0.69
	LPCM + SSE					1	1
	LPCM + ISE						1
0.9	PCM	1	0.91	1	0.90	0.99	1
	MFRM		1	0.92	0.98	0.91	0.92
	MFRM + OIE*			1	0.91	0.99	1
	LPCM				1	0.92	0.92
	LPCM + SSE					1	1
	LPCM + ISE						1

Note: * Data generating model.

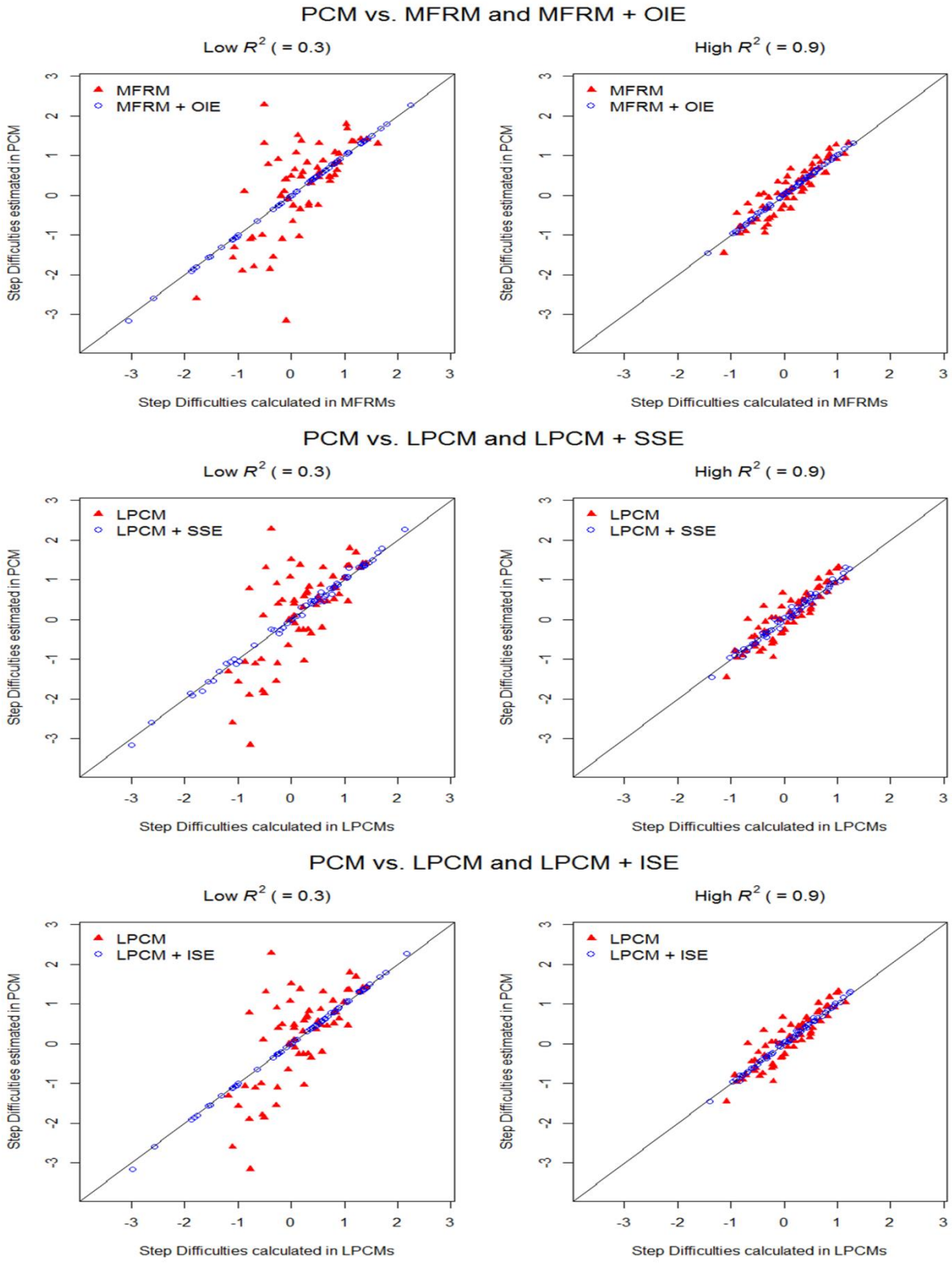


Figure 4-3. Graphical agreement between the models fitted to the MFRM + OIE generation data

These results were confirmed by a graphical comparison of the estimated and calculated step difficulties between the models in Figure 4-3. In all panels, the step difficulty points for the three random item effects models (blue circles) were lying nearly on the 45-degree line indicating perfect alignment. Regarding the two fixed item property effects models (red triangles), compared to the right panels for the high information condition, all the left panels for the low information condition displayed clearly more dispersion of the step difficulty points and they were farther to the 45-degree line, meaning lower agreement.

From the findings in the MFRM approach-based model comparison simulation, in both information conditions, the three polytomous item explanatory models with random item errors performed relatively well in terms of statistical and practical significance. The models without item error terms performed poorly in both information conditions in terms of statistical and practical significance.

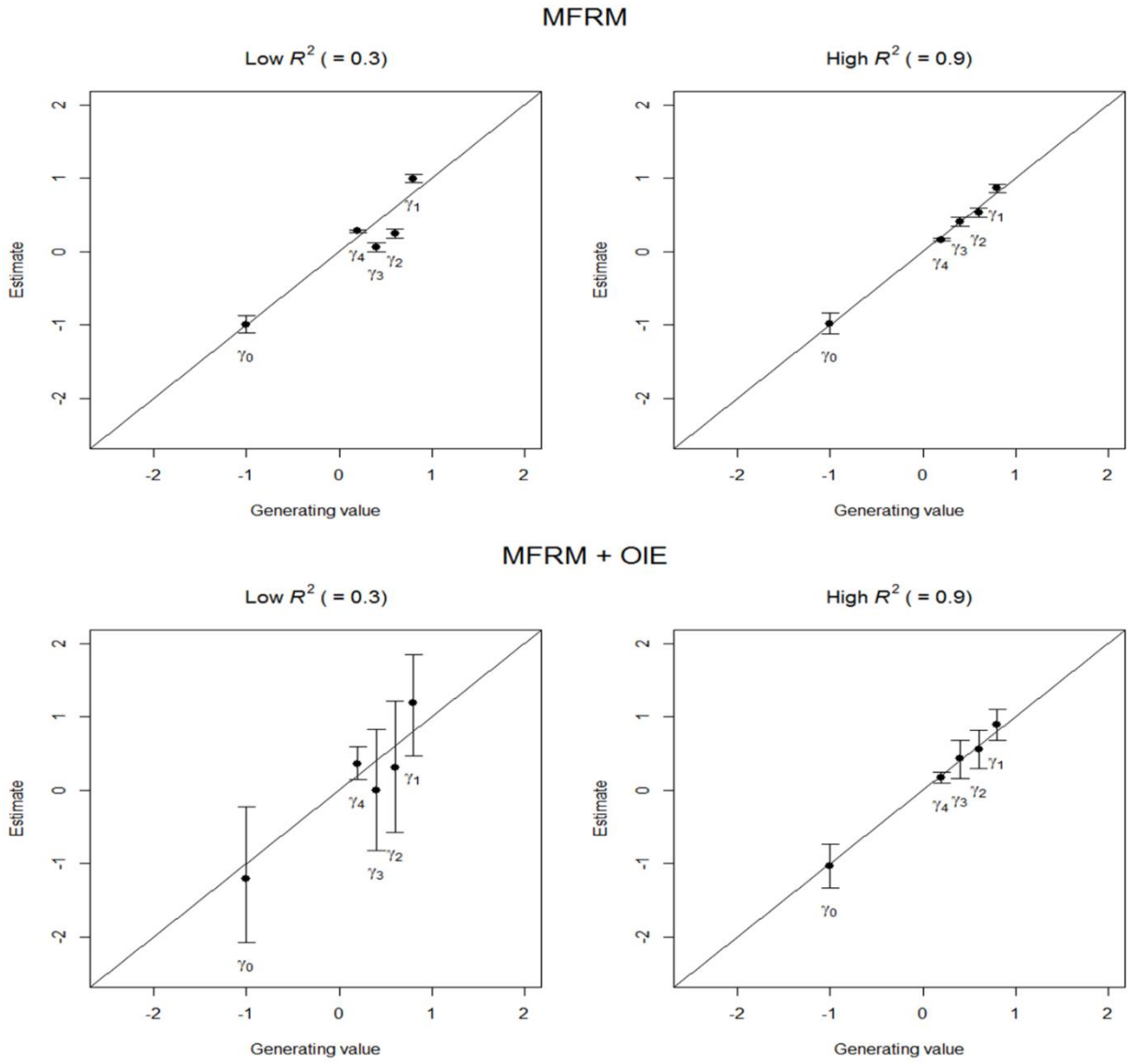


Figure 4-4. Scatter plot with 95% credible intervals for the MFRM + OIE-based simulation

In addition, using a scatter plot of the parameter estimates with 95% credible intervals in Figure 4-4, the effects of model misspecification on the item property effect parameters were examined. In this MFRM approach-based simulation, the data generating model was the MFRM + OIE so that the item property effects on the overall item difficulties compared the MFRM + OIE with a misspecified model, the MFRM.

In the top panels, for some of the parameters, the 95% credible intervals in the MFRM were too narrow to contain the true generating values on the 45-degree line. This phenomenon becomes exacerbated in the low information condition, which suggests that the MFRM is inadequate for making accurate statistical inferences such as estimation and hypothesis testing for the item property effects. In contrast, the MFRM + OIE, correctly specified, in the bottom panels provided sufficiently broad credible intervals to contain the generating values for all the parameters in both information conditions. This suggests that the MFRM + OIE is better for making accurate statistical inferences.

While the MFRM provided narrow credible intervals which represents small standard error estimates, the MFRM with overall item random error provided broad credible intervals which indicates large standard error estimates. These results correspond to other applications of the LLTM + ε to dichotomous data (e.g. De Boeck, 2008; Furr, 2017b; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009).

Linear Partial Credit Model approach-based simulation

In the LPCM approach-based model comparison simulation, the five polytomous item explanatory models with and without random item errors and the PCM were fitted to two sets of the LPCM + SSE generation data with the two information conditions. Regarding convergence, the largest $\hat{R}$ statistic value was 1.03 and the effective sample sizes were large enough in all fitted models.

Table 4-7 presents the goodness-of-fit results of the six models and Figure 4-5 displays them graphically. The order of DIC, LOOIC, and WAIC across the models was consistent except that the PCM and the MFRM + OIE had almost the same values of LOOIC and WAIC in the low information ($R^2 = 0.3$) condition. The data generating model, the LPCM + SSE appeared to be the best fitting model in both information conditions, as found in the MFRM approach-based simulation. It was revealed that the LPCM + ISE and the MFRM + OIE fit worse than the PCM in the low R^2 condition but fit better than the PCM in the high R^2 condition. Putting the findings from the MFRM approach-based simulation together, overall, the goodness-of-fit performance of the LPCM + ISE gets better as explanatory information increases.

The results show that the MFRM and the LPCM fit worse than the PCM in both information conditions, which is as expected because the fixed item property effects models like the LLTM fit worse than the saturated model in general (e.g. Kubinger, 2009; Tuerlinckx & Wang, 2004). Compared to these two polytomous item explanatory models without random item errors, those models with random item errors showed a significantly superior goodness-of-fit in both information conditions. This is because the random item errors account for residual variation in the polytomous item difficulties as in the LLTM + ε applications (e.g. De Boeck, 2008; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009).

Table 4-7. Goodness-of-fit of models fitted to the LPCM + SSE generation data

R^2	Model	k	DIC	LOOIC	WAIC
0.3	PCM	61	22574.51	22589.67	22583.79
	MFRM	36	23541.13	23551.29	23546.00
	MFRM + OIE	37	22577.66	22589.25	22583.52
	LPCM	11	24644.75	24653.18	24648.05
	LPCM + SSE*	14	22572.00	22584.49	22578.87
	LPCM + ISE	12	22582.20	22595.19	22589.36
0.9	PCM	61	24281.60	24294.08	24288.20
	MFRM	36	24404.95	24414.93	24409.44
	MFRM + OIE	37	24281.26	24291.00	24285.15
	LPCM	11	24424.34	24432.78	24427.44
	LPCM + SSE*	14	24272.46	24282.43	24276.69
	LPCM + ISE	12	24277.04	24287.13	24281.29

Note: * Data generating model; k = The number of estimated parameters.

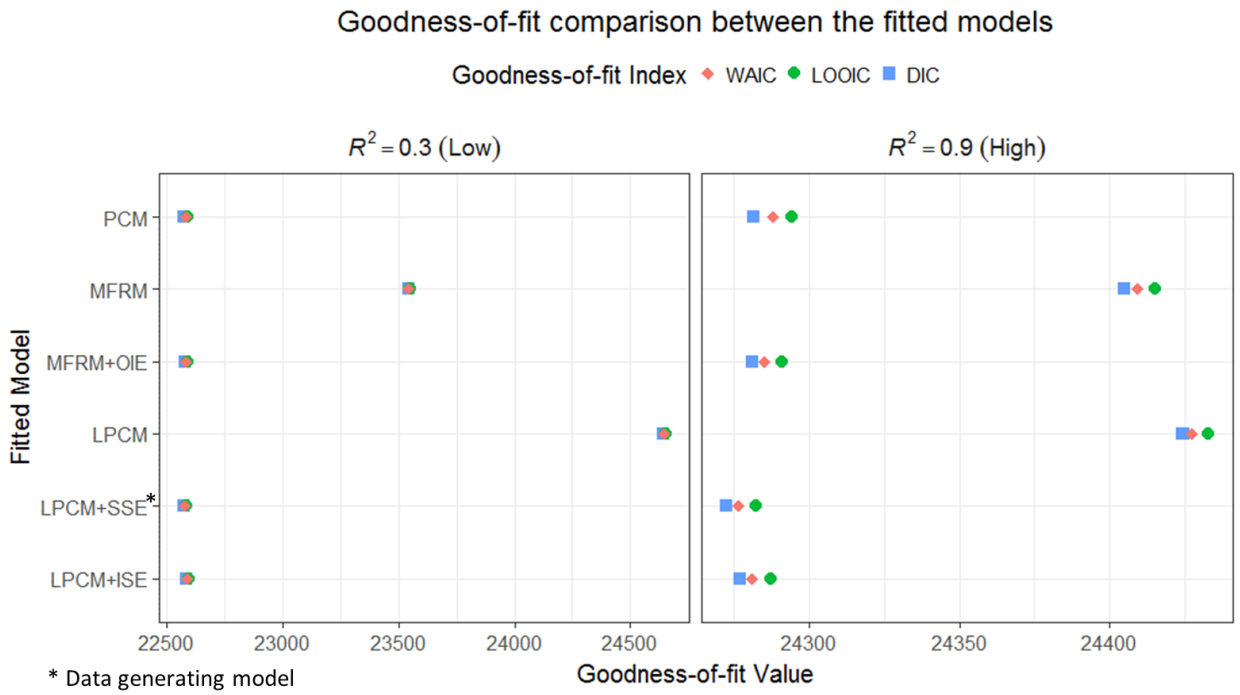


Figure 4-5. Goodness-of-fit of models fitted to the LPCM + SSE generation data

To examine the practical significance of the models, correlations were obtained between the directly estimated step difficulties in the PCM and the calculated step difficulties in the five polytomous item explanatory models, as in Table 4-8. In both information conditions, the three proposed random item effects models revealed perfect agreement with the PCM in that correlations

were 1. This suggests that random item errors worked well to enhance prediction of the overall item difficulties or the step difficulties by allowing for residual variation in both information conditions. The two fixed item property effects models provided similar high agreement in the high information condition (0.99 for the MFRM and 0.98 for the LPCM), whereas the MFRM (0.94) appeared to have better agreement than the LPCM (0.84) in the low information condition. This makes sense because the MFRM has the larger number of parameters than the LPCM even when there is a lack of explanatory information.

Figure 4-6 shows a graphical comparison of the estimated and calculated step difficulties between the models, which confirmed the correlation results. In all panels, the step difficulty points for the three random item effects models (blue circles) were lying nearly on the 45-degree line indicating perfect alignment. Regarding the two fixed item property effects models (red triangles), all the left panels for the low information condition displayed obviously more dispersion of the step difficulty points and they were farther to the 45-degree line, meaning lower agreement, compared to the right panels for the high information condition. Within the left panels, the top panel for the MFRM showed more points close to the line than the other bottom panels for the LPCM.

From the findings in the LPCM approach-based model comparison simulation, in both information conditions, the three polytomous random item effects models performed relatively well in fitting to the LPCM + SSE generation data and performed well in reconstructing the step difficulties. The two fixed item property effects models performed poorly in fitting to the LPCM + SSE generation data in both information conditions and performed differently in reconstructing the step difficulties between the two information conditions.

Table 4-8. Correlations of the estimated and calculated step difficulties between the models fitted to the LPCM + SSE generation data

R^2	Model	PCM	MFRM	MFRM + OIE	LPCM	LPCM + SSE	LPCM + ISE
0.3	PCM	1	0.94	1	0.84	1	1
	MFRM		1	0.94	0.9	0.94	0.94
	MFRM + OIE			1	0.84	1	1
	LPCM				1	0.85	0.85
	LPCM + SSE*					1	1
	LPCM + ISE						1
0.9	PCM	1	0.99	1	0.98	1	1
	MFRM		1	0.99	0.99	0.99	0.99
	MFRM + OIE			1	0.98	1	1
	LPCM				1	0.99	0.99
	LPCM + SSE*					1	1
	LPCM + ISE						1

Note: * Data generating model.

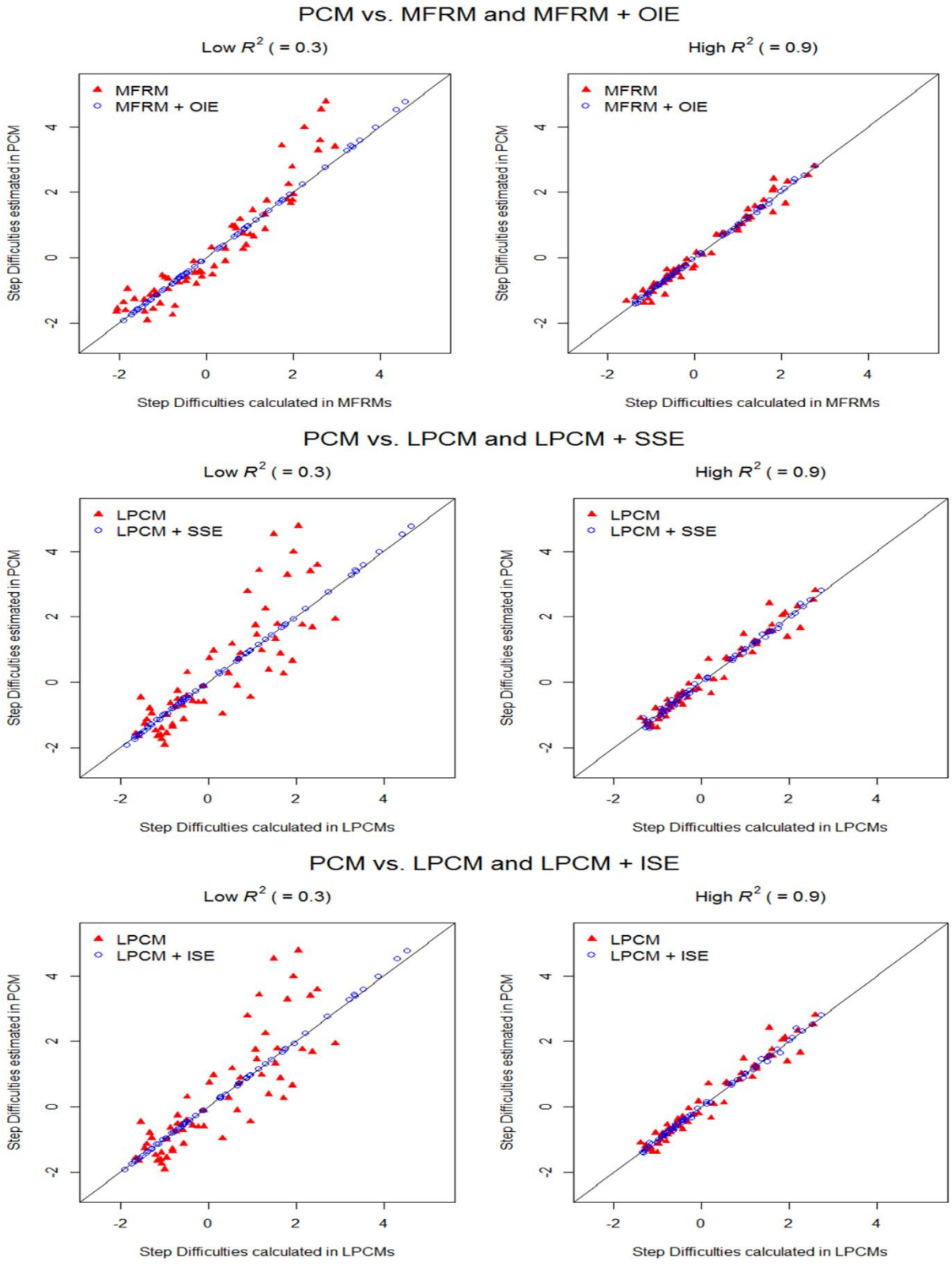


Figure 4-6. Graphical agreement between the models fitted to the LPCM + SSE generation data

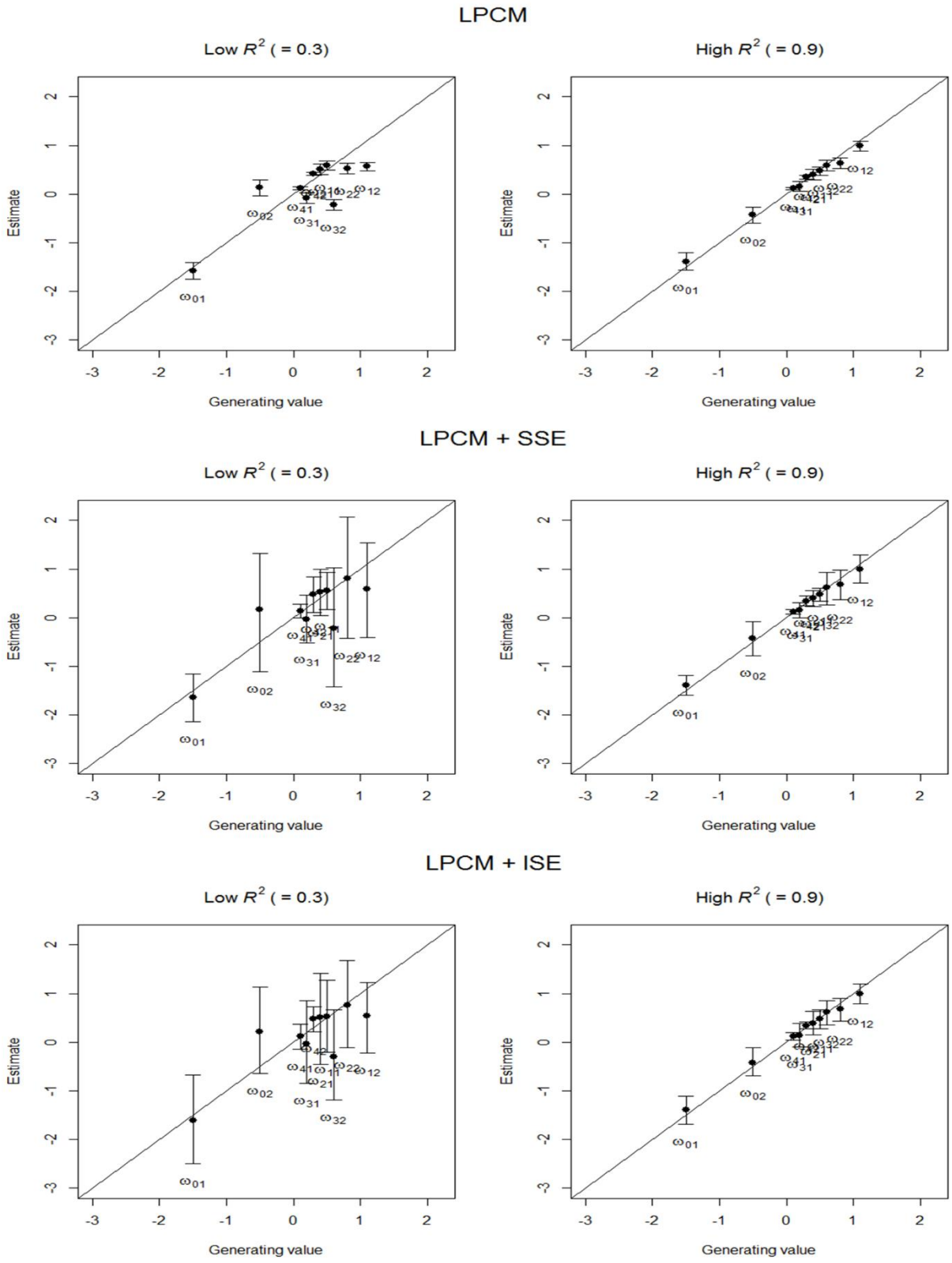


Figure 4-7. Scatter plot with 95% credible intervals for the LPCM + SSE-based simulation

In addition, using a scatter plot of the parameter estimates with 95% credible intervals in Figure 4-7, I examined the effects of model misspecification on the item property effect parameter estimates. In this LPCM approach-based simulation, the data generating model was the LPCM + SSE so that the item property effects on the step difficulties compared the LPCM + SSE with misspecified models, the LPCM and the LPCM + ISE.

In the top panels, the 95% credible intervals in the LPCM were too narrow to contain the true generating values for all the parameters on the 45-degree line. This phenomenon becomes exacerbated in the low information condition, which implies that the LPCM is inadequate for making accurate statistical inferences such as estimation or hypothesis testing for the item property effects. In contrast, the LPCM + SSE, correctly specified, in the middle panels provided sufficiently broad credible intervals to contain the generating values for all the parameters in both information conditions. Despite the misspecified model, the LPCM + ISE in the bottom panels showed comparable parameter estimation with the LPCM + SSE. This suggests that the LPCM + SSE and the LPCM + ISE are better for making accurate statistical inferences.

While the LPCM provided narrow credible intervals which represents small standard error estimates, the other two LPCM with polytomous random item errors provided broad credible intervals which indicates large standard error estimates. These results correspond to other applications of the LLTM + ε to dichotomous data (e.g. De Boeck, 2008; Furr, 2017b; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009).

4.5.3. Study 2: Parameter Recovery

Analysis

In the second simulation study, I conducted a parameter recovery simulation with 30 replications using each polytomous item explanatory approach. Since I have newly proposed three polytomous item explanatory models with random item errors, it would be helpful to assess the recovery of model parameters and the effects of model misspecifications in the low information ($R^2 = 0.3$) and the high information ($R^2 = 0.9$) conditions.

In the MFRM approach-based simulation, one random item effects model (the MFRM + OIE) and one fixed item effects model (the MFRM) were fitted to the two datasets generated for the two information conditions using the MFRM + OIE. In the LPCM approach-based simulation, two random item effects models (the LPCM + SSE and the LPCM + ISE) and one fixed item effects model (the LPCM) were fitted for the two information conditions using the LPCM + SSE. I examined parameter recovery by using the generating model to analyze the simulation datasets, and examined the effects of model misspecification on parameter estimation by using the other models rather than the generating model in each item explanatory approach.

Bias, Monte Carlo error (MCE), and root mean squared error (RMSE) were computed to evaluate the performance of parameter recovery and the effects of model misspecification. Estimates of the bias were calculated by averaging differences between the posterior mean estimates (parameter point estimates) and the generating values across replications. The MCE is a standard error of the posterior mean estimates, regarded as between-simulation variability, which measures the simulation accuracy or uncertainty given a finite sample size of 30 replications (Koehler, Brown, & Haneuse, 2009). Estimates of the MCE were calculated using asymptotic

methods of dividing the standard deviation of the posterior mean estimates for the different generated datasets by $\sqrt{30}$. Using the bias and the MCE, a bias plot with 95% confidence intervals was made in each simulation condition. Estimates of the RMSE were calculated by taking the square root of the mean squared difference between the parameter point estimates and the generating values over replications.

In addition, focusing on the item property effect parameters, I also examined the parameter estimates (the posterior mean) and the empirical standard error (SE) of the estimates (the posterior standard deviation) for the fixed and random item effects models. These were reported as averaged values over the thirty replications. The MCMC simulation settings using Stan and convergence diagnostics were the same as in the model comparison simulation study.

Results

Many-Facet Rasch Model approach-based simulation

In the MFRM approach-based parameter recovery simulation, the MFRM + OIE and the MFRM were fitted for the two information conditions. Regarding convergence, the largest $\hat{R}$ statistic value was 1.07, the largest averaged value over 30 replications was 1.02, and the effective sample sizes were fairly large in all fitted models and information conditions.

Table 4-9 lists the estimated bias, the MCE, and the RMSE in the two models. The MFRM + OIE was correctly specified and the MFRM was misspecified because the datasets were generated using the MFRM + OIE. To evaluate the recovery of model parameters, I examined the MFRM + OIE across the two information conditions first. In the low information condition, the magnitudes of bias were between −0.10 and 0.13 and the RMSE values ranged from 0.05 and 0.35. In the high information condition, the magnitudes of bias were between −0.06 and 0.04 and the RMSE values ranged from 0.03 to 0.15. As explanatory information (R^2) increases, ranges of both the bias magnitudes and the RMSE values decreased significantly. In other words, the parameter estimates become more biased as explanatory information is reduced. Insufficient information aggravates the uncertainty in explanation and such increase of random noise or errors abates accuracy of parameter estimation.

A bias plot with 95% confidence intervals in Figure 4-8 shows these results graphically. The confidence intervals were based on the MCE which reflects sampling variation between simulated replicates. In the top panels for the MFRM + OIE, the 95% confidence intervals for the bias did not contain zero (i.e., significant bias) for four parameters (γ_1, τ_{51}, τ_{161}, τ_{201}) in the low R^2 condition and for two parameters (τ_{91}, τ_{301}) in the high R^2 condition. Compared to the high information condition, parameter estimation of the item property effects in the low information condition was less accurate due to the larger MCE. γ_1 was significantly biased despite the wider confidence intervals.

To evaluate the effects of model misspecification, I examined the MFRM for the two information conditions. From the results in Table 4-9, the magnitudes of bias were between −0.31 and 0.37 and the RMSE values ranged from 0.09 and 0.50 in the low R^2 condition, and the magnitudes of bias were between −0.04 and 0.06 and the RMSE values ranged from 0.02 to 0.15 in the high R^2 condition. The results revealed that the parameter estimates in the low information condition were much more biased than those in the high information condition.

Table 4-9. Parameter recovery for the MFRM + OIE-based simulation

	$R^2 = 0.3$							$R^2 = 0.9$						
		MFRM + OIE			MFRM				MFRM + OIE			MFRM		
	Gen	Bias	Bias.SE	RMSE	Bias	Bias.SE	RMSE	Gen	Bias	Bias.SE	RMSE	Bias	Bias.SE	RMSE
γ_0	-1.0	-0.10	0.06	0.34	0.12	0.05	0.31		0.01	0.03	0.15	0.02	0.03	0.15
γ_1	0.8	0.13	0.06	0.35	-0.07	0.05	0.28		0.01	0.02	0.09	-0.01	0.02	0.08
γ_2	0.6	0.04	0.05	0.29	-0.09	0.05	0.26		-0.03	0.02	0.11	-0.04	0.02	0.11
γ_3	0.4	0.09	0.06	0.33	0.00	0.05	0.27		-0.02	0.02	0.08	-0.02	0.02	0.08
γ_4	0.2	0.00	0.02	0.10	-0.04	0.02	0.09		0.00	0.01	0.03	0.00	0.00	0.02
τ_{11}	-0.1	-0.02	0.02	0.12	0.22	0.05	0.34		0.01	0.02	0.11	0.04	0.02	0.13
τ_{21}	-0.2	0.03	0.03	0.15	0.26	0.05	0.39		0.00	0.02	0.10	0.01	0.02	0.12
τ_{31}	-0.1	0.00	0.02	0.12	0.34	0.06	0.47		0.04	0.03	0.14	0.06	0.03	0.15
τ_{41}	-0.2	-0.02	0.02	0.10	0.28	0.05	0.38		0.02	0.01	0.07	0.04	0.01	0.09
τ_{51}	-0.1	0.03	0.02	0.09	0.27	0.05	0.37		0.01	0.02	0.11	0.03	0.02	0.11
τ_{61}	-0.2	-0.01	0.02	0.10	0.18	0.03	0.23		0.03	0.02	0.13	0.04	0.02	0.14
τ_{71}	-0.1	-0.02	0.02	0.09	0.21	0.02	0.25		0.03	0.02	0.10	0.05	0.02	0.10
τ_{81}	-0.2	0.00	0.02	0.10	0.23	0.03	0.27		-0.01	0.02	0.11	0.01	0.02	0.10
τ_{91}	-0.1	-0.03	0.02	0.13	0.23	0.04	0.33		-0.05	0.02	0.11	-0.03	0.02	0.10
τ_{101}	-0.2	-0.01	0.02	0.13	0.27	0.04	0.35		0.02	0.03	0.14	0.04	0.03	0.15
τ_{111}	-0.1	0.03	0.03	0.14	0.25	0.03	0.29		0.04	0.02	0.13	0.05	0.02	0.12
τ_{121}	-0.2	0.03	0.02	0.11	0.22	0.02	0.26		0.01	0.02	0.09	0.02	0.02	0.09
τ_{131}	-0.1	-0.02	0.02	0.12	0.24	0.03	0.29		0.00	0.02	0.12	0.02	0.02	0.12
τ_{141}	-0.2	0.01	0.02	0.10	0.26	0.03	0.30		0.00	0.02	0.10	0.02	0.02	0.10
τ_{151}	-0.1	-0.02	0.02	0.13	0.22	0.04	0.29		-0.01	0.02	0.10	0.01	0.02	0.11
τ_{161}	-0.2	-0.05	0.02	0.14	0.17	0.03	0.22		-0.01	0.02	0.08	0.01	0.02	0.08
τ_{171}	-0.1	0.01	0.02	0.11	0.27	0.03	0.32		0.01	0.02	0.12	0.02	0.02	0.12
τ_{181}	-0.2	0.03	0.02	0.11	0.37	0.04	0.41		-0.01	0.02	0.09	0.00	0.02	0.09
τ_{191}	-0.1	-0.01	0.02	0.11	0.27	0.05	0.37		-0.02	0.02	0.12	-0.01	0.02	0.12
τ_{201}	-0.2	0.06	0.02	0.14	0.36	0.05	0.44		0.01	0.02	0.11	0.03	0.02	0.12
τ_{211}	-0.1	-0.01	0.02	0.13	0.19	0.03	0.25		0.02	0.02	0.10	0.03	0.02	0.10
τ_{221}	-0.2	0.00	0.02	0.11	0.30	0.05	0.41		-0.02	0.02	0.10	0.00	0.02	0.10
τ_{231}	-0.1	0.02	0.03	0.14	0.28	0.04	0.36		-0.01	0.02	0.11	0.00	0.02	0.12
τ_{241}	-0.2	0.03	0.02	0.12	0.30	0.06	0.45		-0.02	0.02	0.10	-0.01	0.02	0.10
τ_{251}	-0.1	0.04	0.02	0.13	0.37	0.06	0.50		0.04	0.02	0.12	0.06	0.03	0.15
τ_{261}	-0.2	-0.01	0.02	0.11	0.20	0.04	0.30		0.00	0.02	0.11	0.02	0.02	0.12
τ_{271}	-0.1	-0.01	0.03	0.14	0.35	0.05	0.44		0.00	0.03	0.14	0.02	0.03	0.14
τ_{281}	-0.2	0.03	0.02	0.11	0.25	0.03	0.31		-0.02	0.02	0.11	-0.01	0.02	0.11
τ_{291}	-0.1	0.03	0.02	0.13	0.27	0.04	0.35		0.03	0.02	0.13	0.05	0.03	0.14
τ_{301}	-0.2	0.01	0.03	0.14	0.18	0.08	0.46		-0.06	0.02	0.11	-0.04	0.02	0.11
σ_ϵ	0.85	0.04	0.03	0.16				0.19	0.01	0.01	0.05			
σ_θ	1.50	0.00	0.01	0.05	-0.31	0.02	0.32		0.01	0.01	0.05	-0.02	0.01	0.05

Note: Gen = generating value; Bias.SE = Monte Carlo error of bias; RMSE = root mean squared error.

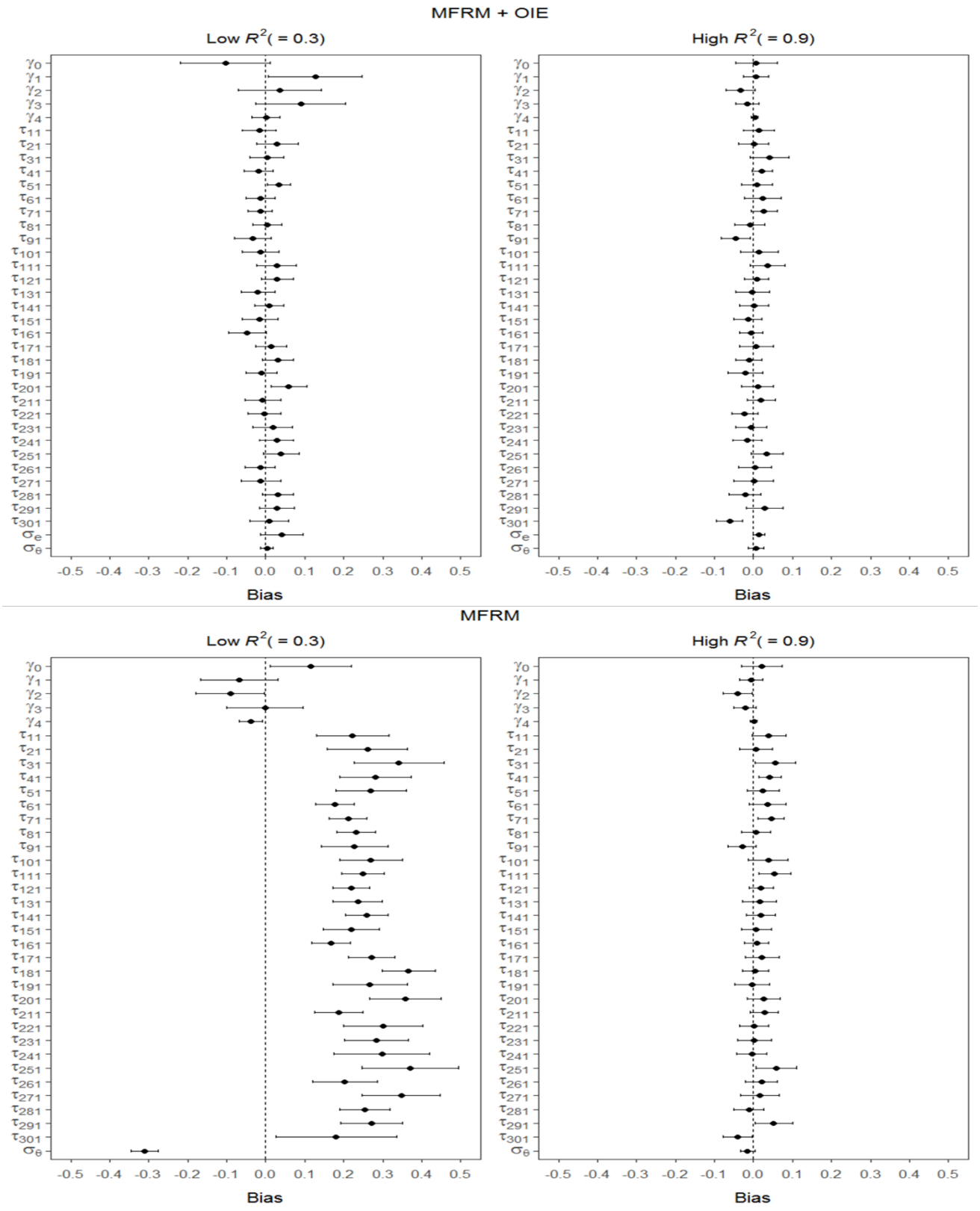


Figure 4-8. Parameter recovery for the MFRM + OIE-based simulation

Compared to the MFRM + OIE, the MFRM gave much larger discrepancies in ranges of both the bias magnitudes and the RMSE values between the two information conditions. In addition, the person standard deviation estimates were biased downwards in both information conditions and the downward bias becomes worsened as R^2 decreases in the MFRM. This indicates shrinkage of the person variance for item explanatory models like the LLTM, which can be understood as a scaling effect that comes from the deficient explanation and prediction of the item difficulty parameters by the fixed item property effects (De Boeck & Wilson, 2004).

These results are displayed clearly using a bias plot with 95% confidence intervals in Figure 4-8. In the bottom panels for the MFRM, the 95% confidence intervals of the parameter point estimates did not contain zero for all parameters except the parameters (γ_1, γ_3) in the low R^2 condition and for eight parameters (γ_2, τ_{31}, τ_{41}, τ_{71}, τ_{111}, τ_{251}, τ_{291}, τ_{301}) in the high R^2 condition. The MFRM provided significantly biased estimates for most of the parameters in the low information condition. When comparing to the top panels for the MFRM + OIE, in both information conditions, more bias points deviated from the zero line for the MFRM. In particular, this phenomenon becomes exacerbated as R^2 decreases, which implies that the MFRM is vulnerable to giving biased parameter estimates when there is poor explanatory information. In contrast, the MFRM + OIE is more robust and accurate for parameter estimation in both information conditions. These results of parameter recovery correspond to an application of the LLTM + ε to dichotomous data by Furr (2017b) although he used different simulation settings.

In addition, focusing on the item property effect parameters, I examined the parameter estimates and their empirical standard error (SE) which were averaged values over replications. Table 4-10 presents the item property effects on the overall item difficulties for the two information conditions.

Table 4-10. Item property effect parameters in the MFRM + OIE-based simulation

		$R^2 = 0.3$				$R^2 = 0.9$			
		MFRM + OIE		MFRM		MFRM + OIE		MFRM	
	Gen	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
γ_0	-1.0	-1.10	0.41	-0.89	0.06	-0.99	0.12	-0.98	0.08
γ_1	0.8	0.93	0.33	0.73	0.03	0.81	0.08	0.79	0.03
γ_2	0.6	0.64	0.41	0.51	0.03	0.57	0.10	0.56	0.03
γ_3	0.4	0.49	0.41	0.40	0.03	0.38	0.10	0.38	0.03
γ_4	0.2	0.20	0.12	0.16	0.01	0.20	0.03	0.20	0.01

Note: Gen = generating value; Estimate = posterior mean; SE = posterior standard deviation (empirical standard error).

In the low R^2 (= 0.3) condition, discrepancy between the generating values and the parameter estimates (i.e., bias) was comparable in that the magnitudes of bias were between −0.10 and 0.13 for the MFRM + OIE and between −0.09 and 0.12 for the MFRM. The standard errors for the MFRM + OIE were much larger than those in the MFRM. In the high R^2 (= 0.9) condition, the magnitudes of bias were between −0.03 and 0.01 for the MFRM + OIE and between −0.04 and 0.02 for the MFRM, which represents comparable bias. The standard error in the MFRM + OIE were larger than those in the MFRM.

In both information conditions, the standard errors of the item property effect parameters were larger for the MFRM + OIE than the MFRM. These results agree with the finding from other LLTM + ε applications (e.g. De Boeck, 2008; Furr, 2017b; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009; Janssen, Schepers, & Peres, 2004). Such small standard errors in the MFRM, the fixed item property effects model, come from omission of the item error reflecting the uncertainty in explanation and prediction.

In addition to the findings from the model comparison simulation, the MFRM is vulnerable for making accurate statistical inference and hypothesis testing for the item property effects particularly with low explanatory information. In contrast, the MFRM + OIE is more robust and accurate for parameter estimation and statistical inference but more rigorous to reject the null hypothesis in statistical testing.

Linear Partial Credit Model approach-based simulation

In the LPCM approach-based parameter recovery simulation, the LPCM + SSE, the LPCM + ISE, and the LPCM were fitted for the two information conditions. Regarding convergence diagnostics, a few $\hat{R}$ statistic values were around 1.1 across 30 replications from the three models, and the largest averaged value over replications was 1.05. The effective sample sizes were acceptably large in all fitted models and information conditions. It was concluded that the MCMC simulations had been converged.

Table 4-11 lists the estimated bias, MCE, and RMSE in the three models. The LPCM + SSE was correctly specified and the LPCM + ISE and the LPCM were misspecified because the datasets were generated using the LPCM + SSE. First, the LPCM + SSE was examined across the two information conditions to evaluate the recovery of model parameters. In the low information ($R^2 = 0.3$) condition, the magnitudes of bias were between −0.06 and 0.09 and the RMSE values ranged from 0.04 and 0.56. In the high information ($R^2 = 0.9$) condition, the magnitudes of bias were between −0.03 and 0.03 and the RMSE values ranged from 0.02 to 0.35. As explanatory information (R^2) increases, ranges of both the bias magnitudes and the RMSE values decreased. As found in the MFRM approach-based simulation, the parameter estimates become more biased as explanatory information is reduced which aggravates the uncertainty in explanation.

In Figure 4-9, a bias plot with 95% confidence intervals displays these results graphically. In the top panels for the LPCM + SSE, e 95% confidence intervals of the parameter point estimates did not contain zero (i.e., significant bias) for the two random item effects parameters ($\sigma_{\xi 1},\ \sigma_{\xi 2}$) in the low R^2 condition and for the random person effects parameter (σ_θ) in the high R^2 condition. Upward bias of the two item error standard deviations makes sense because less information exacerbates random noise or errors. The person standard deviation was a little upwardly biased in the high information condition. This seems that this random item effects model is overfitting. In the model comparison simulation study, the LPCM + SSE showed a superior goodness-of-fit to the PCM and the others in all simulation conditions.

To evaluate the effects of model misspecification, I examined the LPCM + ISE and the LPCM for the two information conditions. From the results in Table 4-11, the LPCM + ISE revealed that the magnitudes of bias were between −0.06 and 0.08 and the RMSE values ranged from 0.04 and 0.55 in the low R^2 condition, and the magnitudes of bias were between −0.04 and 0.03 and the RMSE values ranged from 0.02 to 0.13 in the high R^2 condition. Compared to the

LPCM + SSE, the performance of parameter recovery in the LPCM + ISE was similar although it is misspecified. This means that an item-step random error term in the LPCM + ISE accounts for residual variation in the step difficulties.

Table 4-11. Parameter recovery for the LPCM + SSE-based simulation

		LPCM + SSE			LPCM + ISE			LPCM		
	Gen	Bias	Bias.SE	RMSE	Bias	Bias.SE	RMSE	Bias	Bias.SE	RMSE
$\boldsymbol{R^2 = 0.3}$										
ω_{01}	-1.5	-0.01	0.04	0.20	-0.01	0.04	0.21	0.15	0.04	0.26
ω_{02}	-0.5	0.05	0.10	0.53	0.06	0.10	0.53	0.15	0.09	0.49
ω_{11}	0.5	0.05	0.04	0.21	0.05	0.04	0.21	0.01	0.04	0.20
ω_{12}	1.1	0.09	0.10	0.52	0.08	0.09	0.51	-0.14	0.08	0.46
ω_{21}	0.4	0.02	0.04	0.22	0.02	0.04	0.21	-0.02	0.04	0.22
ω_{22}	0.8	0.05	0.09	0.48	0.05	0.09	0.49	-0.10	0.08	0.46
ω_{31}	0.2	-0.01	0.04	0.20	0.00	0.04	0.20	-0.03	0.04	0.22
ω_{32}	0.6	0.01	0.10	0.56	0.01	0.10	0.55	-0.08	0.09	0.48
ω_{41}	0.1	0.00	0.01	0.07	-0.01	0.01	0.07	-0.01	0.01	0.07
ω_{42}	0.3	-0.06	0.04	0.20	-0.06	0.04	0.20	-0.11	0.03	0.20
$\rho_{\xi 12}$	0.00	0.01	0.04	0.22						
$\sigma_{\xi 1}$	0.51	0.05	0.02	0.11						
$\sigma_{\xi 2}$	1.20	0.07	0.03	0.19						
σ_{θ}	1.50	0.00	0.01	0.04	0.00	0.01	0.04	-0.19	0.01	0.19
$\boldsymbol{R^2 = 0.9}$										
ω_{01}	-1.5	-0.03	0.02	0.12	-0.04	0.02	0.12	-0.03	0.02	0.11
ω_{02}	-0.5	-0.01	0.02	0.12	-0.02	0.02	0.12	-0.01	0.02	0.12
ω_{11}	0.5	-0.01	0.01	0.06	0.00	0.01	0.06	-0.01	0.01	0.06
ω_{12}	1.1	-0.02	0.02	0.11	-0.02	0.02	0.12	-0.03	0.02	0.12
ω_{21}	0.4	-0.01	0.01	0.06	-0.01	0.01	0.07	-0.01	0.01	0.07
ω_{22}	0.8	0.00	0.02	0.10	0.00	0.02	0.10	-0.01	0.02	0.10
ω_{31}	0.2	-0.01	0.01	0.07	-0.01	0.01	0.07	-0.01	0.01	0.07
ω_{32}	0.6	0.01	0.02	0.13	0.01	0.02	0.13	0.00	0.02	0.13
ω_{41}	0.1	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.02
ω_{42}	0.3	0.00	0.01	0.04	0.00	0.01	0.04	0.00	0.01	0.04
$\rho_{\xi 12}$	0.00	-0.03	0.06	0.35						
$\sigma_{\xi 1}$	0.11	0.00	0.01	0.04						
$\sigma_{\xi 2}$	0.26	0.01	0.01	0.04						
σ_{θ}	1.50	0.03	0.01	0.05	0.03	0.01	0.05	0.01	0.01	0.05

Note: Gen = generating value; Bias.SE = Monte Carlo error of bias; RMSE = root mean squared error.

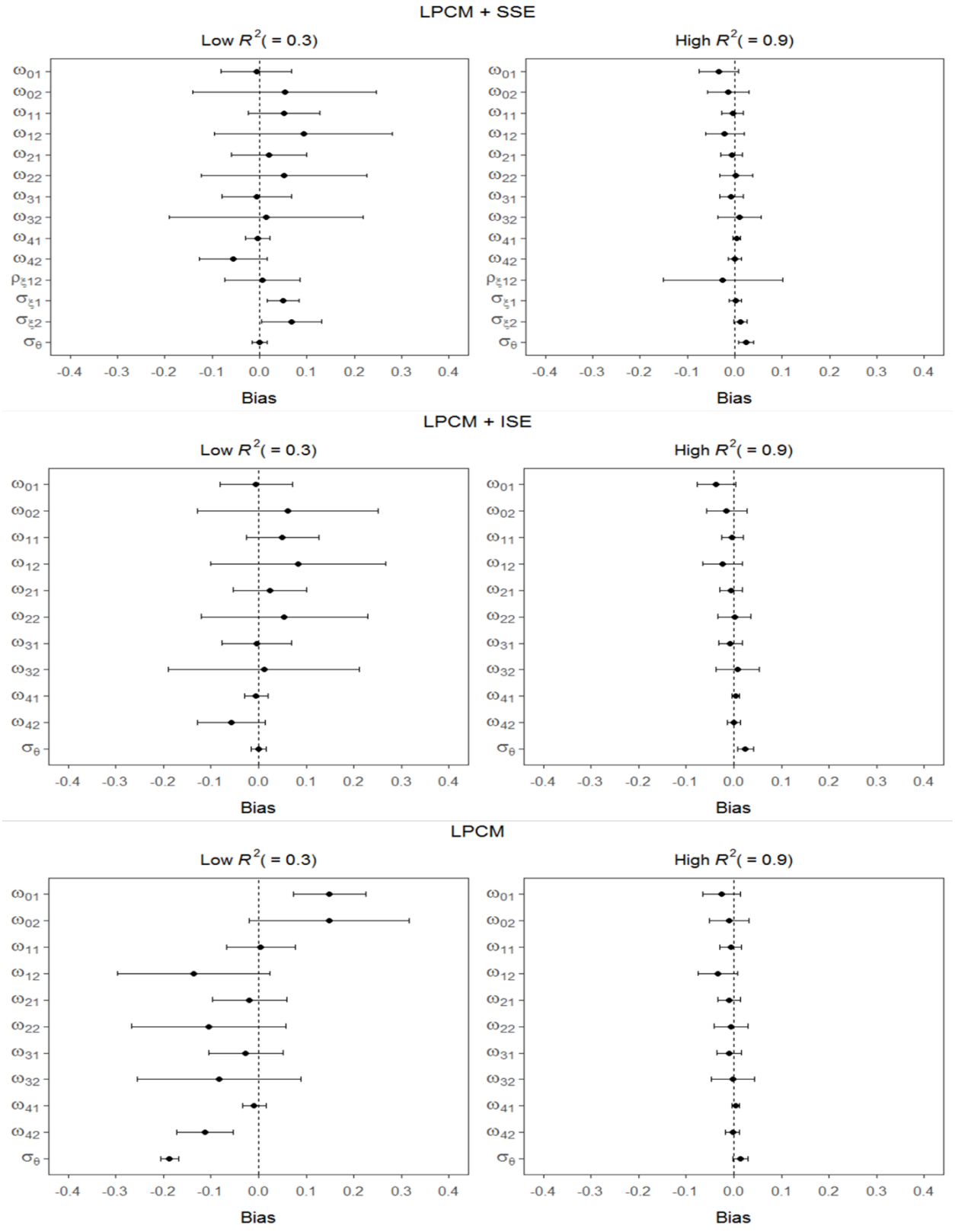


Figure 4-9. Parameter recovery for the LPCM + SSE-based simulation

However, the LPCM revealed larger discrepancies in ranges of both the bias magnitudes and the RMSE values between the two information conditions. In the LPCM, the magnitudes of bias were between −0.19 and 0.15 and the RMSE values ranged from 0.07 and 0.49 in the low R^2 condition, and the magnitudes of bias were between −0.03 and 0.01 and the RMSE values ranged from 0.02 to 0.13 in the high R^2 condition. The parameter estimates in the low information condition were more biased than in the high information condition. In particular, the person standard deviation estimate was biased downwards in the low information because of the scaling effect, as in the MFRM.

These results are displayed in Figure 4-9. In the middle panels for the LPCM + ISE, the 95% confidence intervals of the parameter point estimates contained zero for all parameters in the low R^2 condition but did not for one of the parameters (σ_θ) in the high R^2 condition. In the bottom panels for the LPCM, the 95% confidence intervals of the bias did not contain zero for three parameters ($\omega_{01},\ \omega_{42},\ \sigma_\theta$) in the low R^2 condition and contained zero for all parameters in the high R^2 condition. In the low information condition, in particular, the LPCM provided significantly biased estimates ($\omega_{01},\ \omega_{42},\ \sigma_\theta$), but they were not significantly biased in the other two random item effects models. When there is low explanatory information, the LPCM is vulnerable to giving biased parameter estimates. Overall, the LPCM + SSE and the LPCM + ISE are more robust and accurate for parameter estimation in both information conditions. These results agree with an application of the LLTM + ε to dichotomous data by Furr (2017b) although he used a different simulation setting.

In addition, focusing on the item property effect parameters, I examined the parameter estimates and their empirical standard error (SE) which were averaged values over replications. Table 4-12 presents the step specific item property effects on the overall item difficulties for the two information conditions. In the high R^2 condition, the discrepancies between the generating values and the parameter estimates (i.e., bias) were similar between the three models: the magnitudes of bias were between −0.03 and 0.01 in the LPCM + SSE, between −0.04 and 0.01 in the LPCM + ISE, and between −0.03 and 0.00 in the LPCM. In the low R^2 condition, the bias was similar between the LPCM + SSE and the LPCM + ISE but it was larger in the LPCM: the magnitudes of bias were between −0.06 and 0.09 in the LPCM + SSE, between −0.06 and 0.08 in the LPCM + ISE, and between −0.14 and 0.15 in the LPCM.

In both information conditions, the standard errors of the step specific item property effect parameters were larger for the LPCM + SSE and the LPCM + ISE than the LPCM. These differences were greater in the low R^2 condition. These results agree with the finding from the MFRM approach-based simulation and other LLTM + ε applications (e.g. De Boeck, 2008; Furr, 2017b; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009; Janssen, Schepers, & Peres, 2004). The smaller standard errors in the LPCM comes from omission of the error term which reflects the uncertainty in explanation and prediction.

As found in the model comparison simulation, the LPCM is inadequate for making accurate statistical inferences such as estimation or hypothesis testing for the step specific item property effects in the low information condition. In contrast, the LPCM + SSE and the LPCM + ISE are more robust and accurate for parameter estimation and statistical inferences.

Table 4-12. Item property effect parameters in the LPCM + SSE-based simulation

		LPCM + SSE		LPCM + ISE		LPCM	
	Gen	Estimate	SE	Estimate	SE	Estimate	SE
$\boldsymbol{R^2 = 0.3}$							
ω_{01}	-1.5	-1.51	0.27	-1.51	0.43	-1.35	0.08
ω_{02}	-0.5	-0.45	0.59	-0.44	0.43	-0.35	0.08
ω_{11}	0.5	0.55	0.21	0.55	0.35	0.50	0.05
ω_{12}	1.1	1.19	0.48	1.18	0.35	0.96	0.05
ω_{21}	0.4	0.42	0.26	0.42	0.43	0.38	0.05
ω_{22}	0.8	0.85	0.57	0.85	0.43	0.70	0.06
ω_{31}	0.2	0.19	0.26	0.20	0.43	0.17	0.06
ω_{32}	0.6	0.61	0.57	0.61	0.43	0.52	0.05
ω_{41}	0.1	0.10	0.08	0.09	0.12	0.09	0.02
ω_{42}	0.3	0.24	0.17	0.24	0.12	0.19	0.02
$\boldsymbol{R^2 = 0.9}$							
ω_{01}	-1.5	-1.53	0.11	-1.54	0.13	-1.53	0.09
ω_{02}	-0.5	-0.51	0.15	-0.52	0.13	-0.51	0.09
ω_{11}	0.5	0.50	0.06	0.50	0.09	0.49	0.05
ω_{12}	1.1	1.08	0.11	1.08	0.09	1.07	0.05
ω_{21}	0.4	0.39	0.08	0.39	0.11	0.39	0.06
ω_{22}	0.8	0.80	0.14	0.80	0.11	0.79	0.06
ω_{31}	0.2	0.19	0.08	0.19	0.11	0.19	0.06
ω_{32}	0.6	0.61	0.14	0.61	0.11	0.60	0.06
ω_{41}	0.1	0.10	0.02	0.10	0.03	0.10	0.02
ω_{42}	0.3	0.30	0.04	0.30	0.03	0.30	0.02

Note: Gen = generating value; Estimate = posterior mean; SE = posterior standard deviation (empirical standard error).

4.6. Empirical Studies

For two empirical studies, I used the same data analysis procedure with the model comparison simulation study except for the two information conditions and the effects of model misspecification. Likewise, I used the *RStan* R package (Stan Development Team, 2018) to analyze the two empirical data sets. In addition, to demonstrate practical differences between the MFRM and the LPCM approaches, the effects of item properties on the overall item difficulties or the step difficulties were reported and interpreted in each empirical study.

4.6.1. Application to the Carbon Cycle Assessment data

Data

For the first empirical study, I used a subset of the Carbon Cycle assessment data, pre-test item responses of the Math and Science Partnership (MSP) Carbon student assessment data collected during the 2010-2011 academic year. Middle and high school students from urban, suburban, and rural area in five states (CA, CO, MD, MI, NY) participated in the assessment. The Carbon Cycle assessments were developed for secondary school science education, based on a learning progression for carbon cycling in socio-ecological systems (Jin, Shin, Johnson, Kim, & Anderson, 2015; Mohan, Chen, & Anderson, 2009). Through an iterative process of designing and analyzing the assessments and interviews for students, the learning progression framework (LPF) has been empirically validated in which the research-based construct is defined as four ordered levels of students' understanding and learning about biogeochemical processes in carbon cycling (Jin & Anderson, 2012; Mohan et al., 2009). Under the Carbon Cycle LPF, items were designed and developed to ask students to answer forced-choice questions and explain their choices. The item responses were scored into the four levels of achievement corresponding to students' progress toward more sophisticated reasoning and explanations.

Table 4-13. The Carbon Cycle items and three categorical item properties

Item	Name	Process			Progress				Format	
		Cellular Respiration (CR)	Photosyn-thesis (PS)	Digestion/ Biosynthesis (DB)	Large Scale (LS)	Micro Scale (MS)	Energy (EN)	Mass (MA)	Multiple Choice (MC)	Yes/No Choice (YN)
1	BODYTEMP	1	0	0	0	0	1	0	1	0
2	CARBCYC2	1	0	0	1	0	0	0	0	1
3	CARBPLNT	0	1	0	0	1	0	0	0	1
4	CONTCARB	0	0	1	0	1	0	0	0	1
5	CRKTGRWTH	1	0	0	0	0	0	1	1	0
6	EATAPPLE	0	0	1	0	1	0	0	0	1
7	INFANT	0	0	1	0	0	0	1	0	1
8	LIGHTEN	0	1	0	0	0	1	0	1	0
9	LIGHTEN2	0	1	0	0	0	1	0	1	0
10	MAPLEMASS	0	1	0	0	0	0	1	1	0
11	PLNTGSENS	1	0	0	0	1	0	0	0	1
12	PLNTGRWTH	0	1	0	0	0	0	1	0	1
13	THINGTREE	0	1	0	0	0	0	1	0	1

By the item design, the 13 Carbon Cycle items comprise a combination of three factors of item properties: (a) the Process property has three types of biogeochemical processes in carbon cycling that transform carbon in socio-ecological systems: cellular respiration, photosynthesis, and digestion/biosynthesis, (b) the Progress property has four learning progress variables related to the

carbon cycling process: large-scale systems, micro-scale systems, energy, and mass, and (c) the Format property has two item response formats: multiple choice with explanation, and yes/no choice with explanation. The items are classified by the predictors of each item property, as shown in Table 4-13. In the LLTM approach, these predictors function as weights of the elementary components assembled in a Q matrix (see Kubinger, 2009; Poinstingl, 2009).

Since some items had too few or zero responses for the highest level (e.g. CARBCYC2, CONTCARB, CRKTGRWTH, MAPLEMASS, PLNTGSENS, THINGTREE), all item responses of the four level categories (1, 2, 3, 4) were recoded into three category scores (0, 1, 2, 2) in data analyses to avoid the sparse response problem in item parameter estimation. Hence, the second step's difficulties are interpreted as the added difficulties as one goes from the Level 2 to the combined Levels 3 and 4. Moreover, to improve the quality of estimation and reduce the error of estimation for person abilities, some cases which have less than three valid item responses were removed. In the final sample, in total, 7,139 observations of 1,157 students on the 13 Carbon Cycle items were included for data analyses. In addition, the three categorical item properties were dummy coded; the digestion/biosynthesis in the Process property, the mass in the Progress property, and the yes/no choice with explanation in the Format property served as a reference for each item property.

Results

In the first empirical study, the five polytomous item explanatory models with and without random item errors and the PCM were fitted to the Carbon Cycle assessment data. The $\hat{R}$ statistic values were less than or equal to 1.01 and the effective sample sizes were large enough in all fitted models, suggesting that the MCMC simulations had converged.

Table 4-14. Goodness-of-fit of models fitted to the Carbon Cycle assessment data

Model	*k*	DIC	LOOIC	WAIC
PCM	27	11744.10	11795.72	11765.84
MFRM	21	12605.77	12636.00	12611.46
MFRM + OIE	22	11749.66	11799.33	11768.88
LPCM	15	12816.24	12853.21	12829.93
LPCM + SSE	18	11746.93	11796.99	11766.99
LPCM + ISE	16	11751.70	11803.58	11772.57

Note: *k* = The number of estimated parameters.

Table 4-14 presents the goodness-of-fit results of the six models. The order of DIC, LOOIC, and WAIC across the models was consistent. The PCM, regarded as the saturated model, appeared to be the best fitting model to the Carbon Cycle assessment data. Considering the findings from the model comparison simulation study, this suggests that random item errors in the three proposed models could not fully account for residual variation in the overall item difficulties or the step difficulties of the Carbon Cycle items. Nevertheless, the LPCM + SSE performed fairly well in that values of three goodness-of-fit indices for the LPCM + SSE were close to those for the PCM.

The goodness-of-fit performance of the MFRM and the LPCM was inferior to that of the PCM, which is as expected because fixed item property effects models like the LLTM fit worse than the saturated model in general. The LPCM fit worse than the MFRM because the step deviation parameters were freely estimated for each item in the MFRM and hence it had the larger number of parameters than the LPCM. Compared to these two polytomous item explanatory models without random item errors, those models with random item errors showed a significantly superior goodness-of-fit. This is because random item errors account for residual variation in the polytomous item difficulties as founded in the model comparison simulation study and the LLTM + ε applications (e.g. De Boeck, 2008; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009).

Table 4-15. Correlations of the estimated and calculated step difficulties between the models fitted to the Carbon Cycle assessment data

Model	PCM	MFRM	MFRM + OIE	LPCM	LPCM + SSE	LPCM + ISE
PCM	1	0.91	1	0.91	1	1
MFRM		1	0.91	0.97	0.91	0.92
MFRM + OIE			1	0.91	1	1
LPCM				1	0.91	0.91
LPCM + SSE					1	1
LPCM + ISE						1

To examine practical significance using the agreement between the directly estimated step difficulties in the PCM and the calculated step difficulties in the five polytomous item explanatory models, correlations were obtained (see Table 4-15). The results show that the correlations between the three proposed random item effects models and the PCM were close to 1, meaning perfect agreement. This suggests that the random item errors worked well to enhance prediction of the overall item difficulties or the step difficulties despite the lower goodness-of-fit. The MFRM and the LPCM provided similar agreement (0.91 for both models), although the LPCM fit worse than the MFRM.

As in Figure 4-10, a graphical comparison of the estimated and calculated step difficulties between the models confirmed these results. In three panels, all the step difficulty points for the MFRM + OIE (blue circles), the LPCM + SSE (blue dots), and the LPCM + ISE (green squares) were lying on the 45-degree line indicating perfect alignment. The top panel for the MFRM and the bottom panels for the LPCM displayed similar dispersion of the points. In both models, the first step difficulty of item 13 was located farthest from the 45-degree line.

Recall that there are three categorical item properties (Process, Progress, and Format) in the design of the Carbon Cycle items. To focus on practical differences between the MFRM and the LPCM approaches, the effects of the three item properties on the overall item difficulties or the step difficulties were interpreted in each approach. Table 4-16 shows the results for the fixed and random effects parameters in the two MFRM-based polytomous item explanatory models. The estimated person variance (σ_θ^2) was 0.91^2 for the MFRM and 1.08^2 for the MFRM + OIE. While the MFRM showed shrinkage of the person variance due to the scaling effect that comes from the incomplete explanation of the overall item difficulties by the fixed item property effects (De Boeck,

2008; De Boeck & Wilson, 2004), the MFRM + OIE compensated for the shrinkage by enhancing prediction of the overall item difficulties.

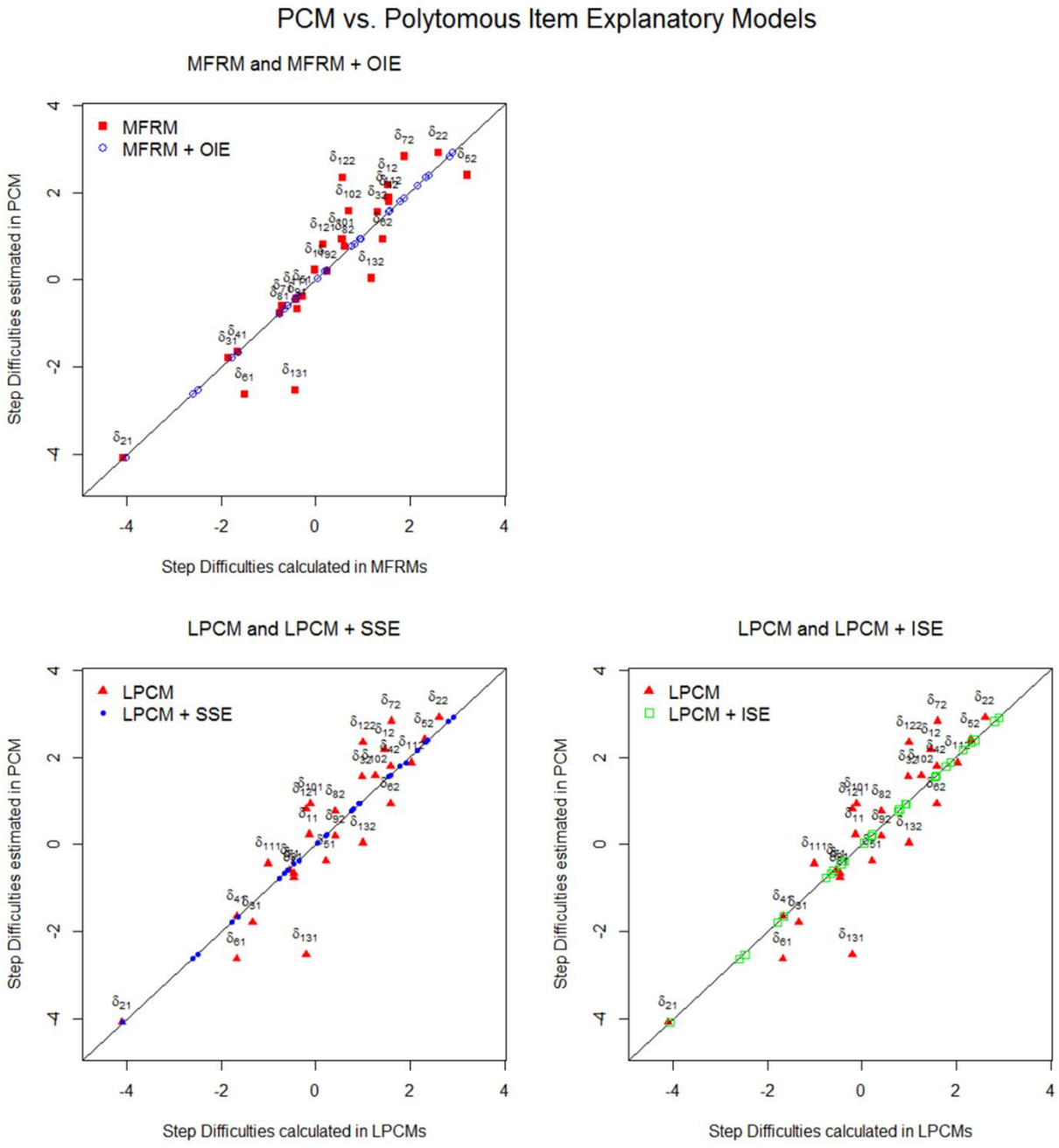


Figure 4-10. Graphical agreement between the models fitted to the Carbon Cycle assessment data

Table 4-16. Model parameter estimates and standard errors in the MFRM-based polytomous item explanatory models to the Carbon Cycle assessment data

Predictor	Parameter	MFRM		MFRM + OIE	
		Estimate	SE	Estimate	SE
Fixed effects					
Intercept	γ_0	0.59	0.07	0.58	1.05
Process (cellular respiration)	γ_{CR}	0.61	0.08	0.61	1.17
Process (photosynthesis)	γ_{PS}	-0.22	0.06	-0.21	1.06
Progress (large-scale systems)	γ_{LS}	-1.94	0.21	-1.75	1.66
Progress (micro-scale systems)	γ_{MS}	-0.64	0.06	-0.72	0.99
Progress (energy)	γ_{EN}	-0.70	0.07	-0.71	1.14
Format (multiple choice)	γ_{MC}	0.26	0.08	0.39	1.22
Step deviation parameters for each item	τ_{11}	-0.78	0.10	-0.96	0.10
	τ_{21}	-3.33	0.19	-3.46	0.20
	τ_{31}	-1.58	0.07	-1.66	0.07
	τ_{41}	-1.60	0.10	-1.72	0.10
	τ_{51}	-1.75	0.11	-1.38	0.10
	τ_{61}	-1.46	0.10	-1.76	0.11
	τ_{71}	-1.30	0.10	-1.71	0.12
	τ_{81}	-0.68	0.10	-0.76	0.10
	τ_{91}	-0.32	0.09	-0.43	0.09
	τ_{101}	-0.07	0.10	-0.32	0.11
	τ_{111}	-0.98	0.10	-1.16	0.10
	τ_{121}	-0.21	0.09	-0.75	0.11
	τ_{131}	-0.81	0.09	-1.27	0.11
Random effects					
Overall item error variance	σ_ϵ^2			1.20^2	0.40
Person variance	σ_θ^2	0.91^2	0.04	1.08^2	0.04

Note: Estimate = posterior mean; SE = posterior standard deviation (empirical standard error); an estimate and a standard error of standard deviation were obtained for the random effects from the Stan output.

In addition, the overall item error variance was estimated as 1.20^2 in the MFRM + OIE. There was significant residual variability in the overall item difficulties in that the item error standard deviation was significant at the 5% level. Based on the estimated item property effects and the overall item error variance, R^2 was calculated as 0.19 using Equation 4.15. This suggests that the three item properties had a low explanatory value on the overall item difficulties of the 13 Carbon Cycle items. In the MFRM + OIE, such low explanatory value is related to the results that all predictors of the item property effects including the intercept were not statistically significant at the 5% level. This is due to the much larger standard errors in the random item effects models than the fixed item property effects models in the low explanatory information (R^2) condition as found in the simulation studies as well as to the small number of items.

Note the MFRM is inadequate for making accurate statistical inferences when there is low explanatory information. Despite the lack of statistical significance, I interpreted the item property effects in the MFRM + OIE to show practical implications and differences between the MFRM and the LPCM approaches (but the step deviation parameters τ are not of interest for interpretation in the MFRM approach).

For the Process property, holding other properties constant, cellular respiration (γ_{CR} = 0.61) increased the overall item difficulty and photosynthesis (γ_{PS} = -0.21) decreased the overall item difficulty compared with digestion/biosynthesis (reference). In the Progress property, keeping other properties constant, compared to mass (reference), large-scale systems (γ_{LS} = -1.75), micro-scale systems (γ_{MS} = -0.72), and energy (γ_{EN} = -0.71) decreased the overall item difficulty, suggesting that large-scale systems was the easiest, then micros-scale systems, followed by energy and finally mass. For the Format property, holding other properties constant, multiple choice with explanation formatted items (γ_{MC} = 0.39) had greater overall item difficulty than yes/no choice with explanation formatted items (reference).

Table 4-17 shows the results for the fixed and random effects parameters in the three LPCM-based polytomous item explanatory models. The estimated person variance (σ_θ^2) was 0.89^2 for the LPCM, 1.09^2 for the LPCM + SSE, and 1.08^2 for the LPCM + ISE. While the LPCM showed shrinkage of the person variance because of the scaling effect, the LPCM + SSE and the LPCM + ISE compensated for the shrinkage by enhancing prediction of the step difficulties. Although the LPCM + ISE performed better than the LPCM in terms of the goodness-of-fit and reconstructing the step difficulties, it fit worse than the LPCM + SSE. Therefore, I focused on the LPCM + SSE in the LPCM approach.

In the LPCM + SSE, the step specific random item error variances were estimated for each step. The item error variance was estimated as 1.37^2 for the first step and 1.02^2 for the second step, and their correlation was 0.60. There was significant residual variability in the step difficulties for each step in that the item error standard deviation for each step was significant at the 5% level. Based on the estimated step specific item property effects and the step specific random item error variances, using Equation 4.15, R^2 was calculated as 0.42 for the first step and 0.31 for the second step. Compared to the MFRM + OIE, R^2 increased in each step, which means that the three item properties had a better explanatory power on the step difficulties for each step. In the LPCM + SSE, two parameters (ω_{02}, ω_{LS1}) appeared to be statistically significant at the 5% level although this model provided the larger standard errors than the LPCM.

Remember that the second step's difficulties represent the relative difficulties as one goes from the Level 2 to the combined Levels of 3 and 4 due to merging the two categories. With regard to the effects of the three item properties on the step difficulties of the first step, keeping other properties constant, cellular respiration process (ω_{CR1} = 0.77, p = 0.555) in the Process property, mass progress variable (reference) in the Progress property, and multiple choice with explanation format (ω_{MC1} = 0.76, p = 0.561) in the Format property made the items more difficult than others within each item property, as a student answered from the Level 1 to the Level 2. In contrast, digestion/biosynthesis process (reference) in the Process property, large-scale systems progress variable (ω_{LS1} = -3.85, p = 0.048) in the Progress property, and yes/no choice with explanation format (reference) in the Format property made the step difficulties less difficult than others within each item property.

For the step difficulties of the second step, holding other properties constant, the cellular respiration process (ω_{CR2} = 0.34, p = 0.728) in the Process property, the large-scale systems

progress variable (ω_{LS2} = 0.32, p = 0.821) in the Progress property, and the yes/no choice with explanation format (reference) in the Format property made it more difficult than others within each item property for a student to achieve either of the Level 3 and Level 4 from the Level 2 in the learning progression for carbon cycling. In contrast, the photosynthesis process (ω_{PS2} = -0.79, p = 0.378) in the Process property, the energy progress variable (ω_{EN2} = -0.77, p = 0.408) in the Progress property, and the multiple choice with explanation format (ω_{MC2} = -0.01, p = 0.993) in the Format property made it less difficult than others within each item property.

Table 4-17. Model parameter estimates and standard errors in the LPCM-based polytomous item explanatory models to the Carbon Cycle assessment data

Predictor	Parameter	LPCM		LPCM + SSE		LPCM + ISE	
		Estimate	SE	Estimate	SE	Estimate	SE
Fixed effects							
Intercept	ω_{01}	-0.55	0.09	-0.99	1.11	-1.08	0.96
	ω_{02}	1.60	0.11	2.25	0.87	2.26	0.97
Process (cellular respiration)	ω_{CR1}	0.67	0.11	0.77	1.30	0.86	1.14
	ω_{CR2}	0.45	0.12	0.34	0.97	0.34	1.16
Process (photosynthesis)	ω_{PS1}	0.34	0.09	0.32	1.10	0.41	0.98
	ω_{PS2}	-0.60	0.10	-0.79	0.90	-0.82	0.98
Progress (large-scale systems)	ω_{LS1}	-4.22	0.37	-3.85	1.95	-3.78	1.61
	ω_{LS2}	0.56	0.23	0.32	1.42	0.28	1.61
Progress (micro-scale systems)	ω_{MS1}	-1.13	0.08	-0.91	1.13	-0.86	0.96
	ω_{MS2}	-0.02	0.09	-0.61	0.83	-0.62	0.95
Progress (energy)	ω_{EN1}	-0.36	0.10	-0.63	1.21	-0.56	1.10
	ω_{EN2}	-0.84	0.13	-0.77	0.93	-0.74	1.09
Format (multiple choice)	ω_{MC1}	0.10	0.10	0.76	1.30	0.68	1.19
	ω_{MC2}	0.26	0.14	-0.01	1.00	-0.07	1.16
Random effects							
Correlation between the steps	$\rho_{\xi 12}$			0.60	0.26		
Item error variance (1st step)	$\sigma^2_{\xi 1}$			1.37^2	0.45		
Item error variance (2nd step)	$\sigma^2_{\xi 2}$			1.02^2	0.36		
Item-Step error variance	σ^2_{ε}					1.16^2	0.28
Person variance	σ^2_{θ}	0.89^2	0.04	1.09^2	0.04	1.08^2	0.04

Note: Estimate = posterior mean; SE = posterior standard deviation (empirical standard error); an estimate and a standard error of standard deviation were obtained for the random effects from the Stan output.

In practice, we could figure out how the three item properties (Process, Progress, and Format) explain the overall item difficulties in the MFRM + OIE as well as the step difficulties in the LPCM + SSE. The two different polytomous item explanatory models demonstrated practical differences in interpreting the effects of the item properties on the overall item difficulties or the

step difficulties of the Carbon Cycle items. Thus, the MFRM and the LPCM approaches are methodologically and practically different item explanatory approaches to polytomous data.

4.6.2. Application to the Verbal Aggression data

Data

In the second empirical study, I used the verbal aggression data set (Vansteelandt, 2000), which is publicly available at the BEAR center website page (see De Boeck & Wilson, 2004; the data set can be downloaded from http://bearcenter.berkeley.edu/EIRM/). The data were collected from the first-year psychology students of a Belgian university. They were asked to answer the behavioral questions about verbally aggressive reactions to frustrating situations. In total, 7,584 observations from 316 persons responding to the 24 verbal aggression items were included in the data set. For polytomous data IRT analyses, the original three ordered-category responses (no = 0, perhaps = 1, yes = 2) were used without a dichotomization.

Influenced by three experimental design factors, the verbal aggression items were designed to have a stem describing a frustrating situation and a verbal aggression response part. The first factor is the Behavior Mode which has two levels of behavior modes—wanting (Want) and doing (Do). For the second factor, the Situation Type has two types of situations—situations in which someone else is to blame (Other-to-blame) and situations in which oneself is to blame (Self-to-blame). The third factor, the Behavior Type, has three kinds of verbal aggressive behaviors—cursing (Curse), scolding (Scold), and shouting (Shout).

For the verbal aggression response part, a total of six (2 × 3) responses were made by mixing the two behavior modes and the three verbal aggressive behaviors. For the frustrating situation, four situations were made: two cases (bus stop, missing a train) for the other-to-blame situation and two cases (grocery store close, disconnected call by an operator) for the self-to-blame situation. A total of 24 (6 × 4) items were written by the item stem which comprises one of the six aggressive responses and one of the four frustrating situations. By the item design, the 24 verbal aggression items are classified by the predictors of each design factor as in Table 4-18.

These item design factors are regarded as categorical item properties and their predictors are functioning as weights of the elementary components gathered into a Q matrix in the LLTM approach. To incorporate the three item properties into polytomous item explanatory models, they were dummy coded; the Want in the Behavior Mode property, the Self-to-blame in the Situation Type property, and the Shout in the Behavior Type property, served a reference for each item property.

Results

For the second empirical study, the five polytomous item explanatory models with and without random item errors and the PCM were fitted to the verbal aggression data. For convergence diagnostics, the MCMC simulations had converged in that the largest $\hat{R}$ statistic value was 1.01 and the effective sample sizes were fairly large in all fitted models.

Table 4-18. The verbal aggression items and three item design factors

Item	Situations	Behavior Mode		Situation Type		Behavior Type		
		Want	Do	Other-to-blame	Self-to-blame	Curse	Scold	Shout
1	A bus fails to stop for me.	1	0	1	0	1	0	0
2		1	0	1	0	0	1	0
3		1	0	1	0	0	0	1
4	I miss a train because the clerk gave me faulty information.	1	0	1	0	1	0	0
5		1	0	1	0	0	1	0
6		1	0	1	0	0	0	1
7	The grocery store closes just as I am about to enter.	1	0	0	1	1	0	0
8		1	0	0	1	0	1	0
9		1	0	0	1	0	0	1
10	The operator disconnects me when I used up my last 10 cents for a call.	1	0	0	1	1	0	0
11		1	0	0	1	0	1	0
12		1	0	0	1	0	0	1
13	A bus fails to stop for me.	0	1	1	0	1	0	0
14		0	1	1	0	0	1	0
15		0	1	1	0	0	0	1
16	I miss a train because the clerk gave me faulty information.	0	1	1	0	1	0	0
17		0	1	1	0	0	1	0
18		0	1	1	0	0	0	1
19	The grocery store closes just as I am about to enter.	0	1	0	1	1	0	0
20		0	1	0	1	0	1	0
21		0	1	0	1	0	0	1
22	The operator disconnects me when I used up my last 10 cents for a call.	0	1	0	1	1	0	0
23		0	1	0	1	0	1	0
24		0	1	0	1	0	0	1

Table 4-19 shows the goodness-of-fit results of the six models. It was revealed that the order of DIC, LOOIC, and WAIC across the models was consistent: the LPCM + SSE, the LPCM + ISE, the MFRM + OIE, the PCM, the MFRM, and the LPCM fit better to the verbal aggression data in sequence. This result agrees with the findings of the high information ($R^2 = 0.9$) condition in the LPCM + SSE-based model comparison simulation. The three random item effects models showed a superior goodness-of-fit to the PCM. This hints that random item errors in the three proposed models could fully account for residual variation in the polytomous item difficulties of the verbal aggression items and/or the three factorial item design worked well for item generation. It is expected that the three item properties (design factors) has a high explanatory value for the verbal aggression items.

The results reveal that the MFRM and the LPCM fit worse than the PCM and the MFRM fit better than the LPCM. These are as expected because, in general, fixed item property effects

models fit worse than the saturated model and the MFRM had the larger number of parameters than the LPCM.

Table 4-19. Goodness-of-fit of models fitted to the verbal aggression data

Model	*k*	DIC	LOOIC	WAIC
PCM	49	12312.49	12324.91	12320.29
MFRM	30	12448.42	12456.71	12452.81
MFRM + OIE	31	12312.39	12322.23	12317.86
LPCM	11	12454.69	12464.35	12460.36
LPCM + SSE	14	12299.66	12308.67	12304.47
LPCM + ISE	12	12307.09	12315.88	12311.65

Note: *k* = The number of estimated parameters.

I looked at practical significance by examining the agreement between the directly estimated step difficulties in the PCM and the calculated step difficulties in the five polytomous item explanatory models via correlations and graphical comparisons. As in Table 4-20, the three random item effects models revealed almost perfect agreement with the PCM in that their correlations were 0.99. This implies that item error terms worked well to enhance prediction of the polytomous item difficulties. The LPCM had a slightly higher correlation with the PCM than the MFRM (ρ = 0.91 for the MFRM, and ρ = 0.93 for the LPCM), although the LPCM fit worse than the MFRM.

Table 4-20. Correlations of the estimated and calculated step difficulties between the models fitted to the verbal aggression data

Model	PCM	MFRM	MFRM + OIE	LPCM	LPCM + SSE	LPCM + ISE
PCM	1	0.91	0.99	0.93	0.99	0.99
MFRM		1	0.94	0.96	0.93	0.94
MFRM + OIE			1	0.94	0.99	1
LPCM				1	0.95	0.96
LPCM + SSE					1	1
LPCM + ISE						1

Figure 4-11 shows a graphical comparison of the estimated and calculated step difficulties between the models, which confirmed the correlation analysis results. In three panels, the step difficulty points for the MFRM + OIE (blue circles), the LPCM + SSE (blue dots), and the LPCM + ISE (green squares) were closer than the two fixed item property effects models to the 45-degree line indicating perfect alignment. Compared to the LPCM in the bottom panels, the MFRM in the top panel had similar dispersion of the points but a few points were farther from the line. The second step difficulty of item 21 were located farthest from the 45-degree line in both models.

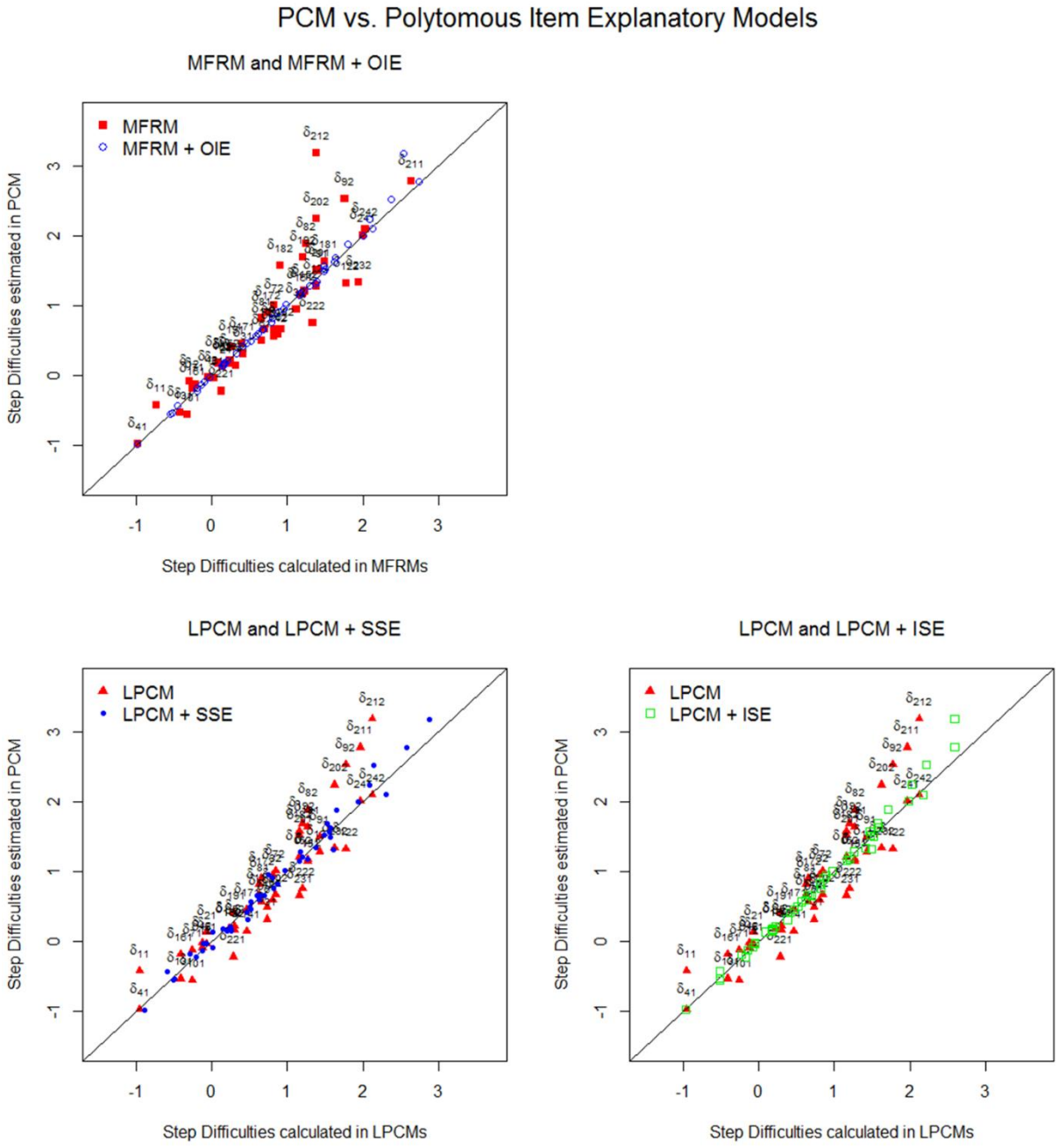


Figure 4-11. Graphical agreement between the models fitted to the verbal aggression data

Remember that the three categorical item properties (Behavior Mode, Situation Type, and Behavior Type) are from the item design factors of the verbal aggression items. Table 4-21 presents the results of the fixed and random effects in the two MFRM-based polytomous item explanatory models. The estimated person variance (σ_θ^2) was 0.95^2 for the MFRM and 0.97^2 for the MFRM + OIE. These similar variance estimates indicate that the shrinkage effect of the person variance estimate was small in the MFRM, whereas the MFRM + OIE compensated for the shrinkage by enhancing prediction of the overall item difficulties.

Table 4-21. Model parameter estimates and standard errors in the MFRM-based polytomous item explanatory models to the verbal aggression data

Predictor	Parameter	MFRM		MFRM + OIE	
		Estimate	SE	Estimate	SE
Fixed effects					
Intercept	γ_0	1.58	0.08	1.69	0.18
Behavior Mode (Do)	γ_{Do}	0.43	0.04	0.49	0.15
Situation Type (Other-to-blame)	γ_{Other}	-0.82	0.04	-0.89	0.14
Behavior Type (Curse)	γ_{Curse}	-1.28	0.05	-1.38	0.17
Behavior Type (Scold)	γ_{Scold}	-0.63	0.05	-0.70	0.18
Step deviation parameters for each item	τ_{11}	-0.22	0.13	-0.17	0.12
	τ_{21}	0.00	0.13	-0.01	0.13
	τ_{31}	-0.35	0.13	-0.32	0.13
	τ_{41}	-0.47	0.12	-0.48	0.13
	τ_{51}	-0.11	0.12	-0.12	0.13
	τ_{61}	-0.11	0.13	-0.05	0.14
	τ_{71}	-0.52	0.12	-0.56	0.12
	τ_{81}	-0.29	0.13	-0.50	0.14
	τ_{91}	-0.18	0.15	-0.45	0.19
	τ_{101}	-0.62	0.12	-0.61	0.11
	τ_{111}	-0.25	0.13	-0.25	0.14
	τ_{121}	-0.19	0.15	-0.04	0.16
	τ_{131}	-0.34	0.12	-0.35	0.12
	τ_{141}	-0.25	0.13	-0.21	0.13
	τ_{151}	-0.03	0.14	-0.03	0.15
	τ_{161}	-0.17	0.13	-0.18	0.13
	τ_{171}	-0.17	0.13	-0.22	0.13
	τ_{181}	0.30	0.15	0.07	0.17
	τ_{191}	-0.48	0.12	-0.62	0.14
	τ_{201}	0.01	0.15	-0.30	0.18
	τ_{211}	0.63	0.21	0.10	0.27
	τ_{221}	-0.60	0.12	-0.49	0.13
	τ_{231}	-0.56	0.13	-0.36	0.13
	τ_{241}	-0.02	0.17	-0.06	0.21
Random effects					
Overall item error variance	σ_ϵ^2			0.32^2	0.07
Person variance	σ_θ^2	0.95^2	0.05	0.97^2	0.05

Note: Estimate = posterior mean; SE = posterior standard deviation (empirical standard error); an estimate and a standard error of standard deviation were obtained for the random effects from the Stan output.

In the MFRM + OIE, the overall item error variance was estimated as 0.32^2, which means significant residual variability in the overall item difficulties. Using Equation 4.15, R^2 was calculated as 0.85 based on the estimated item property effects and the overall item error variance. As expected from the goodness-of-fit results, the three item properties had a high explanatory power on the overall item difficulties of the 24 verbal aggression items.

Although the MFRM + OIE provided larger standard errors than the MFRM, all predictors of the item property effects including the intercept were statistically significant at the 5% level. This was consistent with such a high R^2. The item property effects are interpreted below (but the step deviation parameters τ are not of interest for interpretation in the MFRM approach).

For the Behavior Mode, holding other properties constant, the Doing mode made the overall item difficulty of an item 0.49 logits more difficult than the Wanting mode (γ_{Do} = 0.49, p = 0.001). In the Situation Type, the Other-to-blame situation made the overall item difficulty of an item 0.89 logits easier than the Self-to-blame situation (γ_{Other} = -0.89, p < .001), keeping other properties constant. For the Behavior Type, holding other properties constant, compared to Shouting, Cursing made the overall item difficulty of an item 1.38 logits easier (γ_{Curse} = -1.38, p < .001), and Scolding made the overall item difficulty of an item 0.70 logits easier (γ_{Scold} = -0.70, p < .001).

Table 4-22. Model parameter estimates and standard errors in the LPCM-based polytomous item explanatory models to the verbal aggression data

Predictor	Parameter	LPCM		LPCM + SSE		LPCM + ISE	
		Estimate	SE	Estimate	SE	Estimate	SE
Fixed effects							
Intercept	ω_{01}	1.43	0.08	1.45	0.18	1.45	0.17
	ω_{02}	1.77	0.12	1.96	0.19	1.90	0.20
Behavior Mode (Do)	ω_{Do1}	0.54	0.06	0.53	0.14	0.56	0.14
	ω_{Do2}	0.36	0.07	0.44	0.15	0.42	0.16
Situation Type (Other-to-blame)	ω_{Other1}	-0.70	0.06	-0.71	0.15	-0.70	0.15
	ω_{Other2}	-0.97	0.07	-1.09	0.16	-1.06	0.16
Behavior Type (Curse)	ω_{Curse1}	-1.68	0.08	-1.71	0.17	-1.71	0.19
	ω_{Curse2}	-0.93	0.10	-1.11	0.18	-1.06	0.20
Behavior Type (Scold)	ω_{Scold1}	-0.80	0.07	-0.83	0.19	-0.84	0.18
	ω_{Scold2}	-0.50	0.10	-0.61	0.19	-0.56	0.20
Random effects							
Correlation between the steps	$\rho_{\xi 12}$			0.82	0.16		
Item error variance (1st step)	$\sigma^2_{\xi 1}$			0.33^2	0.07		
Item error variance (2nd step)	$\sigma^2_{\xi 2}$			0.31^2	0.08		
Item-Step error variance	σ^2_{ε}					0.33^2	0.05
Person variance	σ^2_{θ}	0.95^2	0.05	0.97^2	0.05	0.97^2	0.05

Note: Estimate = posterior mean; SE = posterior standard deviation (empirical standard error); an estimate and a standard error of standard deviation were obtained for the random effects from the Stan output.

Table 4-22 shows the results for the fixed and random effects parameters in the three LPCM-based polytomous item explanatory models. The estimated person variance (σ_θ^2) was 0.95^2 for the LPCM, 0.97^2 for the LPCM + SSE, and 0.97^2 for the LPCM + ISE. The shrinkage effect of the person variance estimate was small in the LPCM in that similar variance estimates were across the three models, as found in the MFRM approach. Although the LPCM + ISE performed as well as the LPCM + SSE in reconstructing the step difficulties and in compensating for the person variance shrinkage, the LPCM + SSE performed better in fitting to the verbal aggression data.

From the results of the LPCM + SSE, the item error variance was estimated as 0.33^2 for the first step and 0.31^2 for the second step, and their correlation was estimated as 0.82. Since the item error standard deviation for each step was significant at the 5% level, there was significant residual variability in the step difficulties for each step. Based on the estimated step specific item property effects and the step specific random item error variances, using Equation 4.15, R^2 was calculated as 0.87 for the first step and 0.86 for the second step. These R^2 values were slightly higher than the MFRM + OIE. All predictors of the step specific item property effects including the step intercepts were statistically significant at the 5% level as in the MFRM + OIE, although the LPCM + SSE provided the larger standard errors than the LPCM. This confirmed the high R^2 for both steps.

The effects of the three item properties on the step difficulties of the first step were interpreted as follows. When a person answered "perhaps" rather than "no", keeping other properties constant, the Doing mode made the items 0.53 logits more difficult than the Wanting mode (ω_{Do1} = 0.53, p < .001), the Other-to-blame situation made the items 0.71 logits easier than the Self-to-blame situation (ω_{Other1} = -0.71, p < .001), and compared to Shouting, Cursing made the items 1.71 logits easier (ω_{Curse1} = -1.71, p < .001) and Scolding made the items 0.83 logits easier (ω_{Scold1} = -0.83, p < .001).

For the step difficulties of the second step, holding other properties constant, the Doing mode made the items 0.44 logits more difficult than the Wanting mode (ω_{Do2} = 0.44, p < .001), the Other-to-blame situation made the items 1.09 logits easier than the Self-to-blame situation (ω_{Other2} = -1.09, p < .001), and compared to Shouting, Cursing made the items 1.11 logits easier (ω_{Curse2} = -1.11, p < .001) and Scolding made the items 0.61 logits easier (ω_{Scold2} = -0.61, p < .001), as a person answered "yes" rather than "perhaps".

In practice, we could figure out how the three item design factors (Behavior Mode, Situation Type, and Behavior Type) explain the overall item difficulties in the MFRM + OIE as well as the step difficulties in the LPCM + SSE. The two different polytomous item explanatory approaches demonstrated practical differences in interpreting the effects of the item properties on the overall item difficulties or the step difficulties of the verbal aggression items. Thus, the MFRM and the LPCM approaches are practically different as well as methodologically different.

4.7. Conclusion and Discussion

In this paper, I investigated how to extend the LLTM + ε approach to polytomous data and proposed three polytomous item explanatory models with random item errors using the MFRM and the LPCM approaches: the MFRM + OIE, the LPCM + SSE, and the LPCM + ISE. Using these polytomous item explanatory models, I conducted two simulation studies to investigate

model comparison, parameter recovery, and effects of model misspecification in the low information ($R^2 = 0.3$) and the high information ($R^2 = 0.9$) conditions. I also conducted two empirical studies to demonstrate practical implications and applications of the proposed models to real data sets, the Carbon Cycle assessment data and the verbal aggression data.

From the simulation studies, compared to the two fixed item property effects models, the three proposed polytomous random item effects models performed better in fitting to the simulated data sets as well as in reconstructing the step difficulties in the two information conditions. In each polytomous item explanatory approach, model parameters were recovered better in the high information condition than in the low information condition. The polytomous item explanatory models with random item errors provided larger standard errors of the item property effect parameters than the models without item error terms. By taking the uncertainty in explanation and prediction into account, those random item effects models were more robust and accurate for parameter estimation and statistical inferences.

From the empirical studies, as found in the simulation studies, the three proposed models fit the empirical data better than the two fixed item property effects models. They also performed better in reconstructing the step difficulties. The results showed methodological and practical differences between the MFRM and the LPCM approaches. The MFRM-based item explanatory models explained the overall item difficulties by the item property effects, and overall item random error was allowed for residual variation so that prediction of the overall item difficulties was enhanced. The LPCM-based item explanatory models explained the step difficulties by the step specific item property effects, and step specific random item error or item-step random error was allowed for residual variation so that prediction of the step difficulties was improved.

In educational and psychological measurement research, random item effects models are rather new areas and most applications of them have been limited to dichotomous items (e.g. Cho & Rabe-Hesketh; 2011; De Boeck, 2008; Furr, 2017b; Glas & van der Linden, 2003; Hartig et al., 2012; Holling, Bertling, & Zeuch, 2009; Janssen, Schepers, & Peres, 2004). However, little is known about polytomous items. This paper tried to shed light on the methodological advantages and the practical implications of polytomous random item effects models.

Polytomous random item effects models are emerging and promising. In addition to the findings from the simulation and empirical studies, I emphasize potential uses of them as methodological foundations in IRT modeling. First, although I have focused on item explanatory modeling, i.e., polytomous item explanatory models with random item errors, the proposed models could be combined with other explanatory models such as person explanatory and doubly explanatory models (see De Boeck & Wilson, 2004). Second, if item properties and individual persons interact, the proposed models could be extended to multidimensional versions. This is analogous to the Random-Weights Linear Logistic Test Model (RW-LLTM; Rijmen & De Boeck, 2002) for polytomous items. If item properties and fixed groups of persons interact, the proposed models could be extended with differential facet functioning (DFF; see De Boeck & Wilson, 2004). Third, the proposed models are random item effects models by nature so that they could boost potential uses of dichotomous random IRT models for polytomous items. If the items are nested within item groups, the proposed models could be extended to hierarchical polytomous item structure models with random residuals (see Cho et al., 2014). If finite mixtures of underlying distributions of the person and/or item population are assumed, the proposed models could be combined with latent class models (see De Boeck, 2008; Fieuws, Spiessens, & Draney, 2004; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010). Fourth, the proposed models could be used

for making polytomous item families and item banks. This stresses the generalizability potential and the random sampling interpretation in computer adaptive testing, automatic item generation, and generalizability in item response modeling (see Briggs & Wilson, 2007; Choi, 2013; Holling, Bertling, & Zeuch, 2009; Johnson & Sinharay, 2005; Sinharay, Johnson, & Williamson, 2003).

References

Adams, R. J., & Khoo, S. T. (1996). *Quest*. Melbourne, Australia: Australian Council for Educational Research.

Adams, R. J., & Wilson, M. R. (1996). A random coefficients multinomial logit: A generalized approach to fitting Rasch models. In G. Engelhard & M. Wilson (Eds.), *Objective measurement HI: Theory into practice* (pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of educational and behavioral Statistics*, 22(1), 47-76.

Adams, R.J., Wu, M., & Wilson, M. (2012). *ConQuest 3.0* [computer program]. Hawthorn, Australia: ACER.

Agresti, A. (2002). *Categorical Data Analysis (2nd ed.).* New York: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

Ames, A. J., & Samonte, K. (2015). Using SAS PROC MCMC for item response theory models. *Educational and Psychological Measurement*, 75(4), 585-609.

Andersen, E. B. (1973). Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 26(1), 31-44.

Anderson, J.A. (1984). Regression and ordered categorical variables (with discussion). *Journal of the Royal Statistical Society. Series B*, 46, 1–30.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573.

Atkinson, R. C. & Paulson, J. A. (1972). An approach to the psychology of instruction. *Psychological Bulletin*, 78(1), 49-61.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: A study with normal adults and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241-251. doi: 10.1111/1469-7610.00715.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, 26(3), 349-379.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version, 1.1-17.

Bellio, R., & Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 5(3), 217-227.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5 (1), 7-74.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443-459.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed

models. *Journal of the American Statistical Association*, 88(421), 9-25.
Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4(1), 87-100.
Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, 44(2), 131-155.
Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., & . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1-32.
Cho, S. J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, 55(1), 12-25.
Cho, S. J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: item modeling for explanation and item generation. *Psychometrika*, 79(1), 84-104.
Cho, S. J., Partchev, I., & De Boeck, P. (2012). Parameter estimation of multiple item response profiles model. *British Journal of Mathematical and Statistical Psychology*, 65, 438-466.
Choi, I. H., & Wilson, M. (2015). Multidimensional Classification of Examinees Using the Mixture Random Weights Linear Logistic Test Model. *Educational and Psychological Measurement*, 75(1), 78-101.
Choi, J. I. (2013). *Advances in combining Generalizability Theory and Item Response Theory* (Unpublished doctoral dissertation). University of California, Berkeley.
Clayton, D., & Rasbash, J. (1999). Estimation in large cross random-effect models by data augmentation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(3), 425-436.
Cristante, F., & Robusto, E. (2007). Assessing change with the extended logistic model. *British Journal of Mathematical and Statistical Psychology*, 60(2), 367-375.
Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, 36(1), 1-34.
De Boeck, P. & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer-Verlag.
De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73(4), 533-559.
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.
Draney, K., & Wilson, M. (2008). A LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychology Science Quarterly*, 50(3), 417-432.
Duncan, R. G. & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, 46 (6), 606-609.
Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H).* Strasbourg, France: Council of Europe/Language Policy Division. Retrieved February 18, 2016, from http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf.
Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, 49(2), 175-186.
Embretson, S. E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: LEA.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4), 343-368.
Fieuws, S., Spiessens, B., & Draney, K. (2004). Mixture models. In De Boeck, P. & Wilson, M. (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 317-340). New York: Springer-Verlag.
Fischer, G. H. (1972). A measurement model for the effect of mass-media. *Acta Psychologica*, 36(3), 207-220.
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
Fischer, G. H. (1977). Some probabilistic models for the description of attitudinal and behavioral changes under the influence of mass communication. In W. F. Kempf & B. H. Repp, *Mathematical models for social psychology* (pp. 102-151). Vienna: Hans Huber.
Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In *Handbook of modern item response theory* (pp. 225-243). New York: Springer.
Fischer, G. H., & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika*, 56(4), 637-651.
Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, 59(2), 177-192.
Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.
Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, 47(4), 432-457.
Furr, D. C. (2017a). *Partial credit and generalized partial credit models with latent regression*, Stan Case Studies, Volume 3. http://mc-stan.org/users/documentation/case-studies.
Furr, D. C. (2017b). *Bayesian and frequentist cross-validation methods for explanatory item response models* (Unpublished doctoral dissertation). University of California, Berkeley.
Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.
Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, 54(4), 635-659.
Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, 27(4), 247-261.
Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 505-513.
Haley, S. M., Ni, P., Hambleton, R. K., Slavin, M. D., & Jette, A. M. (2006). Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. *Journal of Clinical Epidemiology*, 59(11), 1174-1182.
Happe, F. G. E. (1995). Understanding minds and metaphors: Insights from the study of figurative language in autism. *Metaphor and Symbolic Activity, 10*(4), 275-295. Doi: 10.1207/s15327868ms1004_3.
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.
Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, 1(2), 81-102.

Hively, W. (1974). Introduction to domain-referenced testing. *Educational Technology*, 14(6), 5-10.

Holling, H., Bertling, J. P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, 35(2-3), 71-76.

Hoskens, M., & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement*, 32(4), 364-384.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In De Boeck, P. & Wilson, M. (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189-212). New York: Springer-Verlag.

Jin, H., & Anderson, C. W. (2012). Developing Assessments for A Learning Progression on Carbon-Transforming Processes in Socio-Ecological Systems. In A. Alonzo & A. Gotwals (Eds.), *Learning Progressions in Science* (pp. 151-181). Rotterdam, Netherlands: Sense Publishers.

Jin, H., Shin, H., Johnson, M. E., Kim, J., & Anderson, C. W. (2015). Developing learning progression-based teacher knowledge measures. *Journal of Research in Science Teaching*, 52(9), 1269-1295.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, 52(12), 5066-5074.

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement*, 29(5), 369-400.

Johnson, T. R. (2007). Discrete choice models for ordinal response variables: A generalization of the stereotype model. *Psychometrika*, 72, 489-504.

Jolin, J. (2015). Using comics to measure social evaluative reasoning in the workplace: Instrument calibration using a Rasch modeling approach. *DADD Online Journal: Research to Practice, 2*(1).

Jolin, J. (2018). *Social Evaluative Reasoning in the Workplace: Validation of an Assessment of Soft Skill Proficiency for Secondary Students in Special Education* (Unpublished doctoral dissertation). University of California, Berkeley.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, 33(7), 499-518.

Kelderman, H., & Rijkes, C. P. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59(2), 149-176.

Koehler, E., Brown, E., & Haneuse, S. J. P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2), 155-162.

Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement*, 69(2), 232-244.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2), 188-229.

Lathrop, Q. N., & Cheng, Y. (2017). Item cloning variation and the impact on the parameters of response models. *Psychometrika*, 82(1), 245-263.

Lei, P. W. (2009). An Overview of the Linear Logistic Test Model. Retrieved March 1, 2013, from https://weblion.psu.edu/trac/warcholak/ticket/33?version=0

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis, 100(9), 1989-2001*.

Linacre, J. M. (1989). *Multi-facet Rasch measurement*. Chicago: MESA Press.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4), 325-337.
Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, 78(3), 384-408.
Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9), 1-20.
Martin, A. D., Quinn, K. M., & Park, J. H. (2011). *Mcmcpack: Markov chain Monte Carlo in R*. *Journal of Statistical Software*, 42(9), 1-21.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, 49(4), 529-544.
Masters, G. N., & Wright, B. D. (1997). The partial credit model. In *Handbook of modern item response theory* (pp. 101-121). New York: Springer.
McCulloch, C.E., Searle, S.R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.
Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19(1), 91-100.
Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, 12(3), 281-296.
Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*, 46(6), 675-698.
Molenaar, I. W. (1983). *Item steps* (Heymans Bulletin 83-630-EX). Groningen, The Netherlands: University of Groningen, Department of Statistics and Measurement Theory.
Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In *Handbook of modern item response theory* (pp. 369-380). New York: Springer.
Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3), 339-348.
Muraki, E., & Bock, R. D. (1997). PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales (computer software). Chicago: Scientific Software International
Pastor, D. A., Dodd, B. G., & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, 26(2), 147-163.
Pearson, P. D. (1978). Questions. In P. D. Pearson & D. D. Johnson (Eds.), *Teaching reading comprehension* (pp. 153-178). New York, NY: Hold, Rinehart, and Winston.
Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the" two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, 24, 307-353.
Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education*, 19(1), 1-20.
Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3), 711-735.
Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), Proceedings of the 3rd

International Workshop on Distributed Statistical Computing (DSC 2003). Retrieved from https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly*, 51(2), 123-134.

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

R Development Core Team. (2008). *R: A language and environment for statistical computing* [computer program]. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual*. UC Berkeley Division of Biostatistics Working Paper Series, Paper No. 160. University of California, Berkeley. (Available from http://www.bepress.com/ucbbiostat/paper160/.)

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., ... & Lewis, T. (2000). A user's guide to MLwiN. London: Institute of Education, 286.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.

Reif, M. (2012). Applying a construction rational to a rule based designed questionnaire using the Rasch model and LLTM. *Psychological Test and Assessment Modeling*, 54(1), 73-89.

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, 26(3), 271-285.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological methods*, 8(2), 185.

Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. van Duijin, & T. Snijders (Eds.), *Essays on item response theory* (pp. 25-42). New York: Springer.

Rost, J., & Langeheine, R. (1997). A guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 13-37). Münster, Germany: Waxmann.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

SAS Institute Inc. (2013). *Getting Started with SAS Enterprise Miner 13*.1. Cary, NC: Author.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

Sinharay, S., Johnson, M.S., & Williamson, D.M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Sciences*, 28, 295–313.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, 50(3), 345-362.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583-639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B*

(Statistical Methodology), 76(3), 485-493.
Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2010). *OpenBUGS Version 3.1.1 user manual*. Retrieved from http://www.openbugs.net/Manuals/Manual.html
Stan Development Team (2017). *Stan Modeling Language Users Guide and Reference Manual*, Version 2.17.0. http://mc-stan.org.
Stan Development Team. (2018). *RStan: the R interface to Stan*. R package version 2.17.3. http://mc-stan.org.
StataCorp. (2015). *Stata Base Reference Manual: Release 14*. College Station, TX: Stata Press.
Tuerlinckx, F., & Wang, W. C. (2004). Models for polytomous data. In De Boeck, P. & Wilson, M. (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75-109). New York: Springer-Verlag.
Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2), 225–255.
Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28(4), 369-386.
Vansteelandt, K. (2000). *Formal models for contextualized personality psychology*. (Unpublished doctoral dissertation), K.U.Leuven, Belgium.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413-1432.
Vehtari, A., Gelman, A., Gabry, J., & Yao, Y. (2018). *loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. R package version 2.0.0.
Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, 11(1), 192-196.
Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594.
Wilson, M. (2005). *Constructing measures: An item response modeling approach.* New York, NY: Taylor & Francis Group.
Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716-730.
Wilson, M., Zheng, X., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, 13(1), 1-22.
Wu, M. L., Adams, R. J., Wilson, M. R., Haldane, S.A. (2007). *ACER ConQuest 2.0 Manual* [computer program]. Hawthorn, Australia: ACER.
Zheng, X., & Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models with gllamm. *The Stata Journal*, 7(3), 313-333.
Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.

Appendices

A.1. Linear Rating Scale Model (LRSM) with a Decomposition of the Item Location Parameters for Chapter 3

Consider an item explanatory extension of the RSM by using the LLTM approach under the MGLMM framework. Note that the item location parameter is interpreted as the overall item difficulty for each polytomous item. In the RSM, item specific properties can be taken into account to explain the overall item difficulties. In a similar way of the MFRM approach to item explanatory models, the item location parameters are imposed to be a linear combination of item specific properties solely, and the scale threshold parameters are freely estimated. Thus, the restricted item location parameters β_i' will now be decomposed into weighted sums of item specific property effect parameters γ_k as follows:

$$\beta_i' = \sum_{k=0}^{K} \gamma_k x_{ik}\,, \qquad k = 0,\ \ldots,\ K \tag{$A1.1$}$$

so that

$$\eta_{pim} = ln\frac{Pr\left(y_{pi} = m\middle|\theta_p\right)}{Pr\left(y_{pi} = m-1\middle|\theta_p\right)} = \theta_p - \sum_{k=0}^{K} \gamma_k x_{ik} - \tau_m, \tag{$A1.2$}$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$, $\eta_{pi0} = 0$, $\tau_0 = 0$, and $\sum_{\mathrm{m}=1}^{M} \tau_m = 0$ ($m = 1,\ \ldots,\ M$). Here, γ_0 is the item intercept representing the difficulty for items with all x_{ik} = 0 for k > 0, γ_k is the regression weight or the effect of item property k on the overall item difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, x_{ik} is the value of item i on item property k, and τ_m is a common scale threshold parameter for the m-th step of all items. For model identification, the mean of the person abilities is constrained to zero. This is the Linear Rating Scale Model (LRSM; Fischer & Parzer, 1991; Fischer, 1997), and also considered as Linacre's (1989) MFRM.

For a simple illustration of the LRSM under the MGLMM framework, I will use the previous example for the MFRM variation. After using dummy coding for the categorical item property (reference for Y), an intercept (γ_0) and the effect of item format X (γ_X) are estimated to account for the overall item difficulties. In addition, one scale threshold parameter (τ_1) is estimated for all items. Note that the linear predictor vector $\boldsymbol{\eta}_p$ contains a cumulative sum of the linear predictor elements η_{pim} of the adjacent-categories logits for all category scores in each item.

First, consider the linear predictor vector for Item 1 which has the X type of item format:
when $m = 0$, $\eta_{p10} = 0$,
when $m = 1$, $\eta_{p10} + \eta_{p11} = 0 + \left(\theta_p - \gamma_0 - \gamma_X - \tau_1\right) = \theta_p - \gamma_0 - \gamma_X - \tau_1$, and
when $m = 2$, $\eta_{p10} + \eta_{p11} + \eta_{p12} = 0 + \left(\theta_p - \gamma_0 - \gamma_X - \tau_1\right) + \left(\theta_p - \gamma_0 - \gamma_X - \tau_2\right) = 2\theta_p - 2\gamma_0 - 2\gamma_X$, due to $\tau_1 + \tau_2 = 0$.

Second, consider the linear predictor vector for Item 2 which has the Y type of item format:

when $m = 0$, $\eta_{p20} = 0$,
when $m = 1$, $\eta_{p20} + \eta_{p21} = 0 + (\theta_p - \gamma_0 - \tau_1) = \theta_p - \gamma_0 - \tau_1$, and
when $m = 2$, $\eta_{p20} + \eta_{p21} + \eta_{p22} = 0 + (\theta_p - \gamma_0 - \tau_1) + (\theta_p - \gamma_0 - \tau_2) = 2\theta_p - 2\gamma_0$, due to $\tau_1 + \tau_2 = 0$.

Therefore, the linear predictor vector for person *p* of the LRSM corresponding to Equation A1.2 can be written as

$$\begin{bmatrix} \eta_{p10} \\ \eta_{p10} + \eta_{p11} \\ \eta_{p10} + \eta_{p11} + \eta_{p12} \\ \eta_{p20} \\ \eta_{p20} + \eta_{p21} \\ \eta_{p20} + \eta_{p21} + \eta_{p22} \end{bmatrix} = \theta_p \times \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 0 & 0 \\ 1 & 0 \\ 2 & 0 \end{bmatrix} \times \begin{bmatrix} \gamma_0 \\ \gamma_X \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \times [\tau_1]. \qquad (A1.3)$$

In this linear predictor vector, the design matrix $\boldsymbol{X}_p$ and the item parameter vector $\boldsymbol{\beta}$ are divided into two parts respectively: the item specific properties (item predictors) and corresponding item specific property effect parameters (γ_0, γ_X), and and the step predictors and corresponding scale threshold parameter (τ_1).

A.2. *gllamm* Syntax for Polytomous Item Explanatory IRT Models for Chapter 3

```
*****************************************************************
** Many-Facet Rasch Model with a Decomposition of the Item Location Parameters
** Item Location Explanatory Partial Credit Model (MFRM approach)
*****************************************************************

*** Load data
insheet using "VA_Long_Polytomous.csv", clear

*** Data Transformation
  ** Make the intercept
gen intcpt=1

  ** Data expansion
tab y               // 3 categories
gen obs = _n
expand 3

  ** Make Category score x (x = 0, 1, 2, ...)
by obs, sort: gen x = _n-1
gen chosen = y == x

*** Make Design Matrix
tab item, gen(i)

  ** Design Matrix using twofold item parameterization
    * Make item location parameter (beta) for each item property
foreach var of varlist intcpt do other curse scold {
 gen B`var' = -1*`var'*x
 }
    * Make step paremater (tau) for each item
forvalues i=1/24 {
  forvalues step=1/2 {
   gen t`i'_`step' = -1*i`i'*(x==`step')
  }
 }
sort id item x

*** Run GLLAMM for Item Location Explanatory Partial Credit Model ***
  ** Should delete Missing values before fitting
drop if y == .
```

```
  ** Run the model
eq slope: x
gllamm x Bintcpt Bdo Bother Bcurse Bscold t*_1, i(id) eqs(slope) link(mlogit) ///
        expand(obs chosen o) noconstant adapt

*****************************************************************
** Linear Partial Credit Model with a Decomposition of the Step Difficulty Parameters
** Step Difficulty Explanatory Partial Credit Model (LPCM approach)
*****************************************************************

*** Load data
insheet using "VA_Long_Polytomous.csv", clear

*** Data Transformation
  ** Make the intercept
gen intcpt=1

  ** Data expansion
tab y               // 3 categories
gen obs = _n
expand 3

  ** Make Category score x (x = 0, 1, 2, ...)
by obs, sort: gen x = _n-1
gen chosen = y == x

*** Make Design Matrix
tab item, gen(i)

  ** Design Matrix using onefold item parameterization
foreach var of varlist intcpt do other curse scold {
  forvalues step=1/2 {
    gen D`var'_`step' = -1*`var'*(x>=`step')
  }
}
sort id item x

*** Run GLLAMM for Step Difficulty Explanatory Partial Credit Model ***
  ** Should delete Missing values before fitting
drop if y == .

  ** Run the model
eq slope: x
gllamm x Dintcpt* Ddo* Dother* Dcurse* Dscold*, i(id) eqs(slope) link(mlogit) ///
        expand(obs chosen o) noconstant adapt
```

A.3. *Stan* Syntax for Polytomous Item Explanatory Models with Random Item Errors for Chapter 4

Many-Facet Rasch Model with Overall Item Random Error (MFRM + OIE)

```
functions {
  real pcm(int y, real theta, vector delta) {
    vector[rows(delta) + 1] unsummed;
    vector[rows(delta) + 1] probs;
    unsummed = append_row(rep_vector(0.0, 1), theta - delta);
    probs = softmax(cumulative_sum(unsummed));
    return categorical_lpmf(y + 1 | probs);
}}
data {
  int<lower=1> P;                    // number of persons
  int<lower=1> I;                    // number of items/questions
  int<lower=1> M;                    // number of steps per item (same for all items)
  int<lower=1> N;                    // number of observations (P*I; as if they were in a vector)
  int<lower=1, upper=P> pp[N];   // person p for observation n
  int<lower=1, upper=I> ii[N];     // item i for observation n
  int<lower=0> y[N];                 // reponse/score for observation n: y = 0, 1 ... m_i
  int<lower=1> K;                    // number of item predictors
  matrix[I,K] Q;                     // matrix of item predictors
}
parameters {
  vector[K] betaip;              // betaip: item property effects on overall item difficulty
  vector[M-1] tau_free[I];      // tau: step deviation
  vector[I] ierror;
  real<lower=0> sigmai;        // sigmai: sd of item errors
  vector[P] theta;               // theta: person ability
  real<lower=0> sigmap;       // sigmap: sd of person abilities
}
transformed parameters {
  vector[I] beta;                // beta: overall item difficulty
  vector[M] tau[I];             // tau: step deviation
  vector[M] delta[I];          // delta: step difficulty
  beta = Q*betaip + ierror;
  for(i in 1:I) for(m in 1:M-1)
    tau[i,m] = tau_free[i,m];
  for(i in 1:I)
    tau[i,M] = -1*sum(tau_free[i,1:M-1]);
  for(i in 1:I) for(m in 1:M)
    delta[i,m] = beta[i] + tau[i,m];
}
```

```
model {                                    // Case constraint (mean of thetas equals to zero)
  target += normal_lpdf(theta | 0, sigmap);
  sigmap ~ cauchy(0, 5) T[0, ];            // hyperprior for theta's sd
  target += normal_lpdf(betaip | 0, 10);
  target += normal_lpdf(ierror | 0, sigmai);
  sigmai ~ cauchy(0, 5) T[0, ];
  for(i in 1:I)
    target += normal_lpdf(tau[i] | 0, 3);
  for (n in 1:N)
    target += pcm(y[n], theta[pp[n]], delta[ii[n]]);    // likelihood
}
generated quantities {
  vector[N] log_lik;
  real deviance;
  for (n in 1:N)
    log_lik[n] = pcm(y[n], theta[pp[n]], delta[ii[n]]);
  deviance = sum(-2*log_lik);
}
```

Linear Partial Credit Model with Step Specific Random Item Error (LPCM + SSE)

```
functions {
  real pcm(int y, real theta, vector delta) {
  vector[rows(delta) + 1] unsummed;
  vector[rows(delta) + 1] probs;
  unsummed = append_row(rep_vector(0.0, 1), theta - delta);
  probs = softmax(cumulative_sum(unsummed));
  return categorical_lpmf(y + 1 | probs);
}}
data {                                  // vectorized version
  int<lower=1> P;                       // number of persons
  int<lower=1> I;                       // number of items/questions
  int<lower=1> M;                       // number of steps per item (same for all items)
  int<lower=1> N;                       // number of observations (P*I; as if they were in a vector)
  int<lower=1, upper=P> pp[N]; // person p for observation n
  int<lower=1, upper=I> ii[N];    // item i for observation n
  int<lower=0> y[N];                    // reponse/score for observation n: y = 0, 1 ... m_i
  int<lower=1> K;                       // number of item predictors
  matrix[I,K] Q;                        // matrix of item predictors
}
parameters {
  matrix[K,M] deltaip;                  // deltaip: item property effects on step difficulty
  matrix[I,M] sserror;
  corr_matrix[M] Omega;
  vector<lower=0>[M] sigmad;    // sd of step difficulties for each step
```

```
  vector[P] theta;                  // theta: person ability
  real<lower=0> sigmap;             // sigmap: sd of person abilities
}
transformed parameters {
  matrix[I,M] delta;                // delta: calculated step difficulty
  cov_matrix[M] Vard;
  delta = Q*deltaip + sserror;
  Vard = quad_form_diag(Omega, sigmad);
}
model {                             // Case constraint (mean of thetas equals to zero)
  matrix[M,M] chol_Sigma;
  target += normal_lpdf(theta | 0, sigmap);
  sigmap ~ cauchy(0, 5) T[0, ];                     // hyperprior for theta's sd
  target += normal_lpdf(to_vector(deltaip) | 0, 10);
  chol_Sigma = cholesky_decompose(Vard);
  target += lkj_corr_lpdf(Omega | 1);
  target += cauchy_lpdf(sigmad | 0, 2.5);
  for (i in 1:I)
    target += multi_normal_cholesky_lpdf(to_vector(sserror[i,]) | rep_vector(0, M), chol_Sigma);
  for (n in 1:N)
    target += pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]]));    // likelihood
}
generated quantities {
  vector[N] log_lik;
  real deviance;
  for (n in 1:N)
    log_lik[n] = pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]]));
  deviance = sum(-2*log_lik);                       //the sum of its unit deviances
}
```

Linear Partial Credit Model with Item-Step Random Error (LPCM + ISE)

```
functions {
  real pcm(int y, real theta, vector delta) {
    vector[rows(delta) + 1] unsummed;
    vector[rows(delta) + 1] probs;
    unsummed = append_row(rep_vector(0.0, 1), theta - delta);
    probs = softmax(cumulative_sum(unsummed));
    return categorical_lpmf(y + 1 | probs);
}}
data {
  int<lower=1> P;                   // number of persons
  int<lower=1> I;                   // number of items/questions
  int<lower=1> M;                   // number of steps per item (same for all items)
  int<lower=1> N;                   // number of observations (P*I; as if they were in a vector)
```

```
  int<lower=1, upper=P> pp[N];  // person p for observation n
  int<lower=1, upper=I> ii[N];     // item i for observation n
  int<lower=0> y[N];                 // reponse/score for observation n: y = 0, 1 ... m_i
  int<lower=1> K;                    // number of item predictors
  matrix[I,K] Q;                     // matrix of item predictors
}
parameters {
  matrix[K,M] deltaip;               // deltaip: item property effects on step difficulty
  matrix[I,M] iserror;
  real<lower=0> sigmais;             // sigmais: sd of item errors
  vector[P] theta;                   // theta: person ability
  real<lower=0> sigmap;              // sigmap: sd of person abilities
}
transformed parameters {
  matrix[I,M] delta;                 // delta: calculated step difficulty
  delta = Q*deltaip + iserror;
}
model {                                          // Case constraint (mean of thetas equals to zero)
  target += normal_lpdf(theta | 0, sigmap);
  sigmap ~ cauchy(0, 5) T[0, ];                  // hyperprior for theta's sd
  target += normal_lpdf(to_vector(deltaip) | 0, 10);
  target += normal_lpdf(to_vector(iserror) | 0, sigmais);
  sigmais ~ cauchy(0, 5) T[0, ];
  for (n in 1:N)
    target += pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]]));    // likelihood
}
generated quantities {
  vector[N] log_lik;
  real deviance;
  for (n in 1:N)
    log_lik[n] = pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]]));
  deviance = sum(-2*log_lik);
}
```