Modeling Rater Effects and Complex Learning Progressions using Item Response Models

By

Hyo Jeong Shin

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 E Hubbard

Spring 2015
Abstract

Modeling Rater Effects and Complex Learning Progressions using Item Response Models

by

Hyo Jeong Shin

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

This dissertation is comprised of three papers that propose and apply psychometric models to deal with complexities and challenges in large-scale assessments, focusing on modeling rater effects and complex learning progressions. In particular, three papers investigate extensions and applications of multilevel and multidimensional item response models, with a primary focus on (1) detecting rater effects in double-scored performance assessments, (2) monitoring human raters with automated scoring engine, and (3) developing measurement models for complicated learning progressions.

The first paper applies and assesses the trifactor model for multiple ratings data in double-scored performance assessments, in which two different raters give independent scores for the same responses (e.g., the GRE essay). The trifactor model incorporates a cross-classified structure (e.g., items and raters) in addition to the general dimension (e.g., examinees). The paper includes a simulation design that follows the GRE example to reflect the incompleteness and imbalance in the real world assessments. The effect of the missingness rate in the data and ignoring the differences among the raters are investigated using the simulations. The use of the trifactor model is illustrated with empirical data.

The second paper applies mixed-effects ordered probit models for the purpose of examining the effectiveness and efficiency of utilizing scores from automated scoring engines (AE) to monitor and provide diagnostic feedback to human raters under training compared to the scores from the human experts (HE). Using the real rater training study data, three types of rater effects—severity, accuracy, and centrality of each rater—are related with model parameters, and compared for cases (a) when the AE is considered as the true score and (b) when the HE is considered as the true score.

The third paper proposes a structured constructs model based on change-point analysis to deal with complicated learning progressions, in which relations between levels across multiple constructs are assumed in advance. Based on the change-point analysis, and reparameterizations of the multidimensional Rasch model and partial credit model, cut score parameters and discontinuity parameters are incorporated to classify the examinees into the levels in the learning progressions, and to model the hypothesized relations as the advantage for examinees belonging to a certain level in one construct to reach a level in another.
construct. Parameter recovery of the proposed model and the consequences of ignoring the hypothesized relations are assessed using simulations. The use of the proposed model is illustrated with empirical data and interpreted as contributing to validity evidence for the hypothesized relations.
Acknowledgements

I wish to express my sincere thanks to the following people whom I owe deepest gratitude. First of all, I would like to thank my primary advisor, Mark Wilson, for his support and valuable guidance. I would also like to thank my secondary advisor, Sophia Rabe-Hesketh, for her kind support and encouragement. My gratitude is also due to Alan Hubbard, for his keen advice on my dissertation. I am also grateful to Karen Draney, for her kind and continuous encouragement over past years.

My gratitude must go out to the Korean Foundation for the Advanced Studies, which has supported me financially for my first five years of graduate studies. Being selected as a scholarship recipient enabled me not only to come to UC Berkeley, but also to enjoy every possible opportunity to learn and grow. I am deeply grateful to my future mentor, Matthias von Davier and Kentaro Yamamoto at Educational Testing Service. I also thank to my colleagues and friends in Quantitative Methods and Evaluation program at UC Berkeley and my dearest friends in Berkeley. Without their support and friendship, I would not survive from this long journey.

Lastly but most importantly, I greatly appreciate my family in New Jersey and South Korea. None of these would have been possible without their endless love and sacrifice. I thank my parents, Hyun Ae Yoon and Joon Dong Shin, who has always believed in and supported me. I also acknowledge my parents in law, Jong Cheol Choi and Kye Ja Hong. I also place on record, my brother’s family, Youngsub Shin, Haelim Choi, and Jay Shin. Most of all, I would like to thank my husband, Sangkook Choi, and my son, Justin Kunhee Choi. Their love and support make me special and my life in Berkeley and NJ blissed.
# Table of Contents

Chapter 1. General Introduction ........................................................................................................ 1

Chapter 2. An Application of the Rater Trifactor Model for the Multiple Ratings Data . 4

2.1. Introduction ................................................................................................................................. 4

2.2. Methods ........................................................................................................................................ 6

  2.2.1. IRT Models for Detecting Rater Effects in the Multiple Ratings Data ....................... 6

  2.2.2. The Rater Trifactor Model (RTM) for the Multiple Ratings Data ............................... 8

  2.2.3. Estimation ......................................................................................................................... 15

2.3. Simulation Study ......................................................................................................................... 16

  2.3.1. Data Generation ................................................................................................................. 16

  2.3.2. Analysis ............................................................................................................................ 19

  2.3.3. Results .............................................................................................................................. 20

2.4. Empirical Example ..................................................................................................................... 27

  2.4.1. Data Source ....................................................................................................................... 27

  2.4.2. Analysis ............................................................................................................................ 28

  2.4.3. Results .............................................................................................................................. 28

2.5. Conclusion and Discussion ....................................................................................................... 31

Chapter 3. Human Rater Monitoring with Automated Scoring Engines .................. 33

3.1. Introduction ................................................................................................................................. 33

3.2. Methods ........................................................................................................................................ 35

  3.2.1. Mixed-effects ordered probit model for rater severity ..................................................... 36

  3.2.2. Mixed-effects ordered probit model for rater severity and rater-specific

  measurement error variances ....................................................................................................... 37

  3.2.3. Types of Rater Effects ........................................................................................................ 38

  3.2.4. Estimation .......................................................................................................................... 39

3.3. Illustration .................................................................................................................................... 40

  3.3.1. Data ................................................................................................................................. 40

  3.3.2. Descriptive Results ......................................................................................................... 41

  3.3.3. Results for the Mixed-effects Ordered Probit Models .................................................... 42
3.4. Conclusion and Discussion ................................................................. 46

Chapter 4. Structured Constructs Models (SCM) based on the Change-point Analysis 49

4.1. Introduction ............................................................................................. 49

4.2. Structured Constructs Models ................................................................. 51
   4.2.1. Previous Statistical Framework of Structured Constructs Models ............... 51
   4.2.2. Structured Constructs Model based on the Change-Point Analysis ................ 53
   4.2.3. Estimation ............................................................................................. 60

4.3. Simulation Study ....................................................................................... 60
   4.3.1. Data Generation .................................................................................. 60
   4.3.2. Analysis ............................................................................................... 61
   4.3.3. Results ................................................................................................. 62

4.4. Empirical Examples .................................................................................. 67
   4.4.1. Data Sources ....................................................................................... 67
   4.4.2. Analysis ............................................................................................... 68
   4.4.3. Results ................................................................................................. 69

4.5. Conclusion and Discussion ....................................................................... 79

Chapter 5. Summary and Conclusion .............................................................. 82

References ...................................................................................................... 84
List of Tables

Table 2-1. Rater bundle model for two raters rating the same piece of student work............... 7
Table 2-2. Allocation of items and raters in the GRE design.............................................. 18
Table 2-3. Bias and RMSE of the Simulation Study by Missingness Rate Design ................. 20
Table 2-4. Bias and RMSE of the Simulation Study by Model Misspecification....................... 23
Table 2-5. Incidence matrix of observed pairs of raters in the empirical data ....................... 27
Table 2-6. Comparison of the model fit statistics................................................................. 28
Table 2-7. Parameter estimates and standard errors for the Carbon Cycle project data ........ 29
Table 3-1. Four types of analyses.......................................................................................... 36
Table 3-2. Direct comparison between the AE and the HE .................................................. 41
Table 4-1. Bias and RMSE in the baseline models and the change-point SCM ...................... 63
Table 4-2. Parameter estimates and standard errors for the Two Level Case ADM............... 70
Table 4-3. Parameter estimates and standard errors for the data of the three levels of ADM . 73
Table 4-4. Parameter estimates and standard errors for the three levels CK-PCK ............... 77
List of Figures

Figure 2-1. The RTM for multiple ratings data ($I=3, R=2$) ................................................................. 10
Figure 2-2. Connected relationships between the raters under the GRE design ......................... 19
Figure 2-3. Bias and RMSE of the Simulation Study by Missing Rate Design ......................... 22
Figure 2-4. Bias and RMSE of the Simulation Study by Misspecification ............................... 25
Figure 2-5. Distribution of the Median Odds Ratio by Items ............................................................... 26
Figure 2-6. Distribution of the Median Odds Ratio by Raters ......................................................... 26
Figure 3-1. Data Structure .................................................................................................................. 40
Figure 3-2. Severity/leniency of each HR depending on the target score ............................... 44
Figure 3-3. Rater-specific measurement error variance each HR depending on the target score .................................................................................................................................. 45
Figure 4-1. An example of level-specific relationships in the ADM project ............................ 48
Figure 4-2. Hypothesized links between two constructs for the Two level Case and the Three Level Case .................................................................................................................................. 50
Figure 4-3. An example of change-point SCM for the Two Level Case ........................................ 56
Figure 4-4. An example of change-point SCM for the Three Level Case ..................................... 59
Figure 4-5. Estimated model parameters for each cut-score (Two level Case) ....................... 66
Figure 4-6. Estimated model parameters for each cut-score (Three level Case) ....................... 67
Figure 4-7. Comparison of the item parameter estimates ............................................................... 71
Figure 4-8. Distribution of bivariate person proficiencies ................................................................. 72
Figure 4-9. Comparison of the item parameter estimates ................................................................. 75
Figure 4-10. Distribution of bivariate person proficiencies .............................................................. 76
Figure 4-11. Comparison of the item parameter estimates .............................................................. 78
Figure 4-12. Distribution of bivariate person proficiencies .............................................................. 79
Chapter 1.
General Introduction

Constructed response (CR) items have been used widely in many large-scale assessments, such as the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA) and the Graduate Record Examinations (GRE). In particular, CR items are believed to be best suited to test reasoning abilities, higher-level cognitive skills such as problem solving or critical thinking. The use of CR items will be much more increasing with the assessments based on the Common Core State Standards (CCSS), because the CCSS calls for a system that elicits complex student demonstrations or applications of knowledge and skills (Porter, McMaken, Hwang, & Yang, 2011). When CR items are used, responses on CR items should be evaluated by raters, whether by human raters or an automated scoring engine. Each scoring method, human raters and automated scoring engine, addresses different challenges and measurement problems (Zhang, 2013). Human raters can make holistic decisions based on the developed scoring rubrics and their background knowledge. They can evaluate critical thinking skills, including the quality and factual correctness of the responses. However, some criticize this scoring method as expensive, and that human raters can be inconsistent and subjective introducing systematic error from different raters or rater drift over time. As an alternative scoring method, automated scoring has become increasingly used, particularly in the low-stakes settings. These automated systems work algorithmically and are expected to be efficient, consistent, and achieve greater objectivity than human scoring. However, the current state-of-the-art is not sufficiently mature yet to be used as the sole score in the high-stakes decisions. Furthermore, for automated scoring, human ratings are quite often used as a development target as well as evaluation criteria.

Another trend in large-scale assessments is that a learning progression framework is increasingly used as a guiding framework for the Common Core State Standards and Next Generation Assessment Systems, which are intended to bring coherence to curricula, instruction and assessment. Learning progressions (LPs) refer to “descriptions of successively more sophisticated ways of thinking about an important domain of knowledge and practice” (Corcoran, Mosher, & Rogat, 2009). It is essential for researchers to develop and apply the appropriate measurement models to validate the LP empirically and to describe student achievement in terms of levels in the LP, which leads to learning progression level-referenced inferences. However, in large-scale assessments, it may be too simplistic to assume a complete ordering of the ways of thinking for a unidimensional domain, and complex LP may involve a multidimensional structure, each of which has its own ordering, and the sub dimensions may have complex relationships with one another, or links from one level of one dimension to a level of another dimension (Wilson, 2012).

This dissertation consists of three papers that respond to those complexities and challenges in large-scale assessments, focusing on modeling rater effects and complex learning progressions. In particular, three papers investigate extensions and applications of
multilevel and multidimensional item response models, with a primary focus on (1) detecting rater effects in double-scored performance assessments, (2) monitoring human raters with automated scoring engine, and (3) developing measurement models for complicated learning progressions. These three papers correspond to Chapters 2, 3, and 4 respectively in this dissertation and below I provide brief introductions of each chapter.

Chapter 2. An Application of the Rater Trifactor Model for the Multiple Ratings Data

Under the multiple rating situations in which different raters give independent scores for the same responses (e.g., the GRE essay exam), the conditional independence assumption in the conventional IRT models is violated. This leads to the problem of overestimated reliability and underestimated measurement error. As a new approach that corrects the problem and can be more widely and conveniently used, the Rater Trifactor Model (RTM) is described and investigated. The RTM incorporates a cross-classified structure (e.g., items and raters) in addition to the general dimension (e.g., examinees). Estimation is performed using Bayesian methods of Markov chain Monte Carlo (MCMC). Simulation studies are conducted to assess the impacts of missingness rate due to the incompleteness and imbalance in the real world assessments. The consequences of ignoring differences among the raters, particularly in rater-specific measurement variances, are also investigated using the simulations. The use of the RTM is illustrated with empirical data.

Chapter 3. Human Rater Monitoring with Automated Scoring Engines

Interest in applying automated scoring to supplement or supplant human scoring has increased in recent years. This paper aims to evaluate the effectiveness and efficiency of utilizing scores from automated scoring engines (AE) to monitor and provide diagnostic feedback to human raters under training compared to the scores from the human experts (HE). Two mixed-effects ordered probit models, one for rater severity and another for rater severity and rater accuracy, are applied using maximum likelihood estimation for the real rater training study data. Three types of rater effects—severity, accuracy, and centrality of each rater—are related with the model parameters, and compared for cases (a) when the AE is considered as the true score and (b) when the HE is considered as the true score.

Chapter 4. Structured Constructs Models (SCM) based on the Change-point Analysis

Structured constructs models (SCM) are measurement models designed to deal with complicated learning progressions, in which relations between levels across multiple constructs are assumed in advance. This study proposes a SCM formulation based on change-
point analysis, which can be useful in estimating the cut scores and the discontinuities. Based on the reparameterizations of the multidimensional Rasch model and partial credit model, cut score parameters are estimated to classify examinees into the levels on the latent continuum. Further, discontinuity parameters are incorporated to model the hypothesized relations as the advantage for examinees belonging to a certain level in one construct to reach a level in another construct. This study presents the conceptual framework of the proposed SCM for the Two Level Case (e.g., mastery vs. non-mastery) as well as the Three Level Case (e.g., Level 1, Level 2, and Level 3), and estimation based on the MCMC algorithm. Parameter recovery of the proposed model and the consequences of ignoring the hypothesized relations are assessed using simulations. The use of the proposed model is illustrated with empirical data and interpreted as contributing to validity evidence for the hypothesized relations.
Chapter 2.
An Application of the Rater Trifactor Model for the Multiple Ratings Data

2.1. Introduction

Understanding rater effects has been emphasized since constructed responses and open-ended questions have become widely used in performance assessments as well as in large-scale assessments. Unlike multiple-choice items, the constructed responses must be evaluated by human raters or by an automated scoring engine. In the scoring process, typically, different raters evaluate examinees’ responses. Moreover, to improve accuracy of the scoring process, multiple rating scores are often used for the same response, such as for the GRE essay exam. For example in the GRE essay exam, each essay is given scores from two different raters and then averaged scores between two raters are reported to examinees (Educational Testing Service, 2011). Sometimes, when the two scores have large discrepancies, an adjudicated score is additionally given by a third rater. Although many testing programs adopt thorough training process and employ qualified raters, the act of rating is naturally subject to various error and bias sources between raters, which threatens the reliability and validity of the assessment. To address these problems, there have been various efforts to investigate or to reduce rater effects by modeling the measurement error due to raters.

One of the most common and traditional approaches is to examine the inter-rater agreement or inter-rater reliability. Quite often, inter-rater agreement is calculated as the agreement rates between ratings made by different raters, and the indices are easily used as a part of the reliability check in practice. Another approach follows a generalizability theory (G-theory), which provides a way of partitioning the total variance in a set of ratings into separate, uncorrelated parts that are each associated with a different source of variability (Shavelson, Baxter, & Gao, 1993). G-theory can be used to partition the total variability in the ratings into variance components due to each of the sources of variation, such as (a) systematic variability between individual examinees, (b) variability between raters (inter-rater inconsistencies), (c) variability within raters across rating occasions (intra-rater inconsistencies), and (d) variability between the writing tasks (e.g., Sudweeks, Reeve, & Bradshaw, 2004). Unlike these two approaches, which consider only the observed scores, measurement models based on item response theory (IRT) or factor analysis models postulate a latent variable structure based on the observed discrete response data. One of the most popular models within the IRT context is the many-facet Rasch model (Linacre, 1994; Linacre & Wright, 2002), which includes a parameter for rater severity as one of the facets in the model, in addition to the persons and items. This approach has been widely applied to many studies in educational assessments, such as writing assessment (e.g., Engelhard, 2002). From the factor analysis literature, models developed for the data having different raters can be considered as the adaptations of the models for multitrait-multimethod (MTMM).
(Achenbach, 2011). Quite often, the “methods” are raters, and the goal is to isolate latent trait from rater effects and measurement error. Unlike the IRT approach, which has usually focused on estimating individual rater effects and figuring out which rater seems problematic, the primary concern of the models based on the MTMM approach has been to assess convergent and discriminant validity and to obtain the aggregated scores for the trait of examinees.

When the data contains multiple ratings such as GRE essay exam, the problem becomes more complicated due to the violation of the underlying assumptions in the IRT measurement models. In particular, under the multiple rating situations in which different raters give independent scores for the same responses (e.g., essay), the conditional independence assumption in the conventional IRT models is violated. Conditional independence means that within any group of examinees all characterized by the same values (i.e., latent trait), the conditional distributions of the item scores are all independent of each other (Lord, Novick, & Birnbaum, 1968). In other words, the conditional probability of the whole score vector is the product of the conditional probabilities of each item score. Under this assumption, latent person ability is the only source of dependence or correlation between scores. Thus, if there are any other potential sources playing a role in addition to the characterization by the IRT models, it would not be appropriate to assume conditional independence (Embretson & Reise, 2000). In multiple ratings data, there are two types of conditional dependencies, given the latent person ability. First, dependency occurs between ratings given by the same rater. For example, the scores given by one lenient rater are likely to be correlated due to the rater’s characteristic of being lenient. Another dependency comes from the scores assigned to a single response. Since the respondent does not produce each response or essay for each rater, two scores given by different raters share the item being asked, and those scores will be correlated by nature. If ignored, the dependency between ratings results in underestimated measurement error and overestimated reliability (Bock, Brennan, & Muraki, 2002; Verhelst & Verstralen, 2001; Wilson & Hoskens, 2001).

To correct the problem, the rater bundle model (RBM; Wilson & Hoskens, 2001), the hierarchical rater model (HRM; Patz, Junker, Johnson, & Mariano, 2002), and the HRM modification based on signal detection theory (HRM-SDT; DeCarlo, Kim, & Johnson, 2011) have been proposed. In particular, the RBM incorporates an interaction parameter, which models the possibility that raters would agree more for the same item response than we would expect for different raters for different items. The RBM has a unique strength in the explicit parameterization of the dependency using the interaction parameters, but it may be difficult to formulate when there are many different combinations of rater pairs. In contrast, the HRM and the HRM-SDT break the data generation process down into two levels: in the first level, the observed scores given by the raters are related to the “true latent category” by signal detection theory, and in the second level, the probability of being in the true latent category is specified by the IRT models. The HRM and the HRM-SDT are useful for understanding how raters behave in the rating process, but these models assume that the observed score and the true latent score are discrete variables. However, instead of assuming the discrete variable during the rating process, it is plausible to assume that raters imagine a certain continuum when they give scores. As Bejar (2012) noted, a major approach for understanding what goes
on in the minds of raters as they rate can be represented by similarities or distances, which are conceptually closer to a continuum rather than discrete categories.

Thus, although the previous IRT models have their own unique strengths and weaknesses, we identify the need for a new approach that can be more widely and conveniently used in practice. In this regard, the Rater Trifactor Model (RTM) is introduced in this paper, and is investigated as an alternative measurement model for multiple rating data. In contrast to the RBM, the RTM parameterizes rater effects as fixed-effects for the location and random-effects for the variability, which allows us to avoid increasing the number of the model parameters as the number of rater pairs increase. In contrast to the HRM and the HRM-SDT, the RTM assumes that the underlying random variables associated with raters is a continuum, instead of assuming the discrete ordinal variables. In brief, the RTM incorporates a cross-classified structure (e.g., items and raters) in addition to the general dimension (e.g., examinees). This model stipulates that multiple ratings reflect three sources of variability, namely, the examinees, items, and raters.

The goals of this study are twofold. One is to introduce the RTM and demonstrate an application of it to analyzing the multiple ratings data. The other is to investigate the impacts of missingness rate and the impacts of misspecified homogeneity assumption on variances of the latent variables. To this end, this chapter is organized as follows. First, the IRT models for detecting rater effects in the context of multiple ratings data are reviewed and the RTM for the multiple ratings data is introduced. Second, a brief explanation is given as to how the Bayesian methods of Markov chain Monte Carlo (MCMC) can be employed to fit the RTM. Third, two sets of simulation studies are carried out to assess the parameter recovery depending on the missingness rate and misspecification of homogeneity assumption for variances. Fourth, the use of the RTM is illustrated with an empirical example of real data from the Carbon Cycle project (Jin & Anderson, 2012). Lastly, the chapter ends with concluding remarks and suggestions for further studies.

2.2. Methods

2.2.1. IRT Models for Detecting Rater Effects in the Multiple Ratings Data

*Rater effects* can be defined simply as “patterns of ratings that contain measurement error from the raters” (Edward W. Wolfe & McVay, 2012). Raters may introduce errors into examinee scores in various ways—unfamiliarity with or inadequate training in the use of the rating scale, fatigue or lapses in attention, deficiencies in some areas of content knowledge that are relevant to making scoring decisions, or personal beliefs that conflict with the values espoused by the scoring rubric (Myford & Wolfe, 2003, 2004; Saal, Downey, & Lahey, 1980).

Among the many approaches to estimate rater effects, within the IRT literature, the many-facet Rasch model (MFRM; Eckes, 2009; Linacre & Wright, 2002; Linacre, 1994) is a popular choice. In fact, the MFRM has the same mathematical form as the linear logistic test model (LLTM; Fischer, 1973): the MFRM incorporates rater severity parameters into the
LLTM and assumes conditional independence between the ratings. For example, based on the Rasch model (Rasch, 1960), the MFRM can be written as

\[
\text{logit}(\Pr(X_{irp} = 1 | \theta_p)) = \theta_p - \beta_i - \rho_r
\]  

(1.1)

where \(X_{irp}\) is the dichotomous rating given to examinee \(p\) on item \(i\) by rater \(r\), \(\theta_p\) is the latent proficiency of examinee \(p\), \(\beta_i\) is the difficulty of item \(i\), and \(\rho_r\) is the severity of rater \(r\).

Within the context of multiple ratings, the main concern is how to deal with the conditional dependencies among the ratings in an appropriate way. First, Wilson and Hoskens (2001) asserted that ignoring the dependence between ratings in the MFRM leads to a distortion in standard error calculations for the estimates of \(\theta_p\) and other model parameters. As an alternative, they proposed the rater bundle model (RBM), which accounts for the conditional dependencies that occur when raters provide repeated ratings on the same piece of work. They introduced an interaction parameter \((\kappa_{i12})\), which models the possibility that raters will agree more than we would expect if we just knew their rater harshness. The interaction parameters can be parameterized in three ways: (a) an item specific interaction that is constrained to be constant within items regardless of the assigned rater pair \((\kappa_{i12} = \kappa_{..})\), (b) a rater-pair specific interaction that is constrained to be the same for a particular pair of raters, no matter which items \((\kappa_{i12} = \kappa_{..} \text{ for all } i)\), and (c) a single uniform interaction effect across items and across rater pairs \((\kappa_{i12} = \kappa_{..})\). As a simple example for the dichotomous scores, the RBM’s log odds are represented in Table 2-1.

Table 2-1. Rater bundle model for two raters rating the same piece of student work

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Log odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>(-\kappa_{i12})</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>(\theta_p - \beta_i - \rho_{r1})</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>(\theta_p - \beta_i - \rho_{r2})</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>(2\theta_p - 2\beta_i - \rho_{r1} - \rho_{r2} - \kappa_{i12})</td>
</tr>
</tbody>
</table>

Note. \(\rho_1\) is the rater severity parameter for the first rater among two raters, and \(\rho_2\) is the corresponding value for the second rater among two raters.

A comparison between the RBM and the MFRM revealed that reliability is overestimated when conditional rater dependence is ignored in the MFRM. Even more, a comparison of the item difficulties and rater severities estimated by the MFRM and the RBM showed a clear pattern that statistically significant rater severities in the RBM were masked in the MFRM by the misspecification in the model (Wilson & Hoskens, 2001).

Similar to the RBM, the hierarchical rater model (HRM; Patz, Junker, Johnson, & Mariano, 2002; Patz, 1996) and its variations (DeCarlo, Kim, & Johnson, 2011b; Mariano & Junker, 2007) were developed to model conditional dependence in multiple ratings data. In particular, the HRM breaks the data generation process down into two stages. The first stage
is for latent ideal rating variables, which describe examinee \( p \)’s performance on item \( i \) as an unobserved per-item latent discrete variable \( \xi_{pi} \). This ideal rating captures dependence between observed multiple ratings of the same piece of examinee work, and is assumed to follow an IRT model. The second stage is to model the observed ratings, rating \( k \) for examinee \( p \)’s performance on item \( i \), which may be different from the ideal rating category. This is modeled as a discrete signal detection model, using a matrix of rating probabilities that rater \( r \) rates \( k \) given the ideal rating. In most HRM studies, these probabilities were assumed to be proportional to a normal density with a mean of \( \xi + \rho_r \) and standard deviation \( \psi \). To our knowledge, applying the HRM to real or simulated dichotomous data has not been published yet. Although the normal assumption looks inappropriate for dichotomously scored data, the HRM based on the Rasch model can be written as follows:

\[
\begin{align*}
\theta_p &\sim N(\mu, \sigma^2), \\
\text{logit}(\Pr(\xi_{pi} = \xi | \theta_p)) &= \theta_p - \beta_i, \\
\Pr(X_{pir} = k | \xi_{pi} = \xi, X_{pir} \in \{0, 1\}) &\propto \exp\left\{-\frac{1}{2\psi^2_r} (k - (\xi + \rho_r))^2\right\}.
\end{align*}
\]

A comparison between the MFRM and the HRM showed that the HRM fits both the simulated and real data, which are polytomously scored, significantly better when the data presents a dependence structure between multiple ratings (Patz et al., 2002). The HRM was also more effective and unbiased in item parameter recovery than the MFRM. In addition, compared with the HRM, the MFRM produced person proficiency estimates with noticeably underestimated interval widths.

### 2.2.2. The Rater Trifactor Model (RTM) for the Multiple Ratings Data

#### Model Formulations

The IRT models reviewed above have been designed to deal with conditional dependence in the multiple ratings data. Similarly, bifactor item response models have been developed to accommodate conditional dependence in a different situation (Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992; Jeon, Rabe-Hesketh, & Rijmen, 2013; Wang, Bradlow, & Wainer, 2002), such as so-called testlets or item bundles, in which items share a common stimulus (e.g., Rosenbaum, 1988; Wainer & Kiely, 1987). The bifactor model for the dichotomous score using the logit link is written as

\[
\text{logit}(\Pr(X_{pir} = 1 | \theta_p)) = \alpha_i \theta_p + \alpha_{it} \theta_{pt(i)}^T - \beta_i
\]

where \( \alpha_{it} \) is the loading or discrimination parameter for item \( i \) for the \( t \)th specific dimension \( \theta_{pt(i)}^T \) that item \( i \) belongs to, \( \alpha_i \) is the loading for item \( i \) for the general dimension \( \theta_p \), and \( \beta_i \)
is the location for item $i$. The latent variables are all assumed to be independent of each other and follow standard normal distributions, $\theta_p \sim N(0,1)$ and $\theta_p \sim N(0,1)$. Instead of constraining the latent variable distributions to the standard normal distributions, we can constrain the loadings of the items alternatively. In particular, when the loadings of all items are constrained to be 1, the model corresponds to the multidimensional Rasch model with zero correlations between the latent variables. By constraining the discrimination parameters, the variances of the latent variables can then be freely estimated.

de la Torre and Song (2009) proposed a second-order IRT model, which can be understood as the discrete response analog of a second-order factor model. As Rijmen (2010) showed, the second-order model is equivalent to the testlet model, which imposes a random variable for each testlet, and have proved to be special cases of bifactor models. Recently, Rijmen, Jeon, von Davier, and Rabe-Hesketh (2014) extended the model of de la Torre and Song to a third-order model. In their study, the second-order model has been extended to accommodate a higher-order structure of the latent variables, which can be also called a trifactor model, where multiple topic areas are embedded within each multiple-content domain (Rijmen et al., 2014; Rijmen, 2011). They also showed how both the second- and third-order structures allow for efficient full-information maximum likelihood estimation. A similar model can be found in Bauer et al., (2013) from factor analysis framework. In their study, the primary goal was to evaluate and measure a single construct for the individuals based on the ratings from multiple informants (e.g., parents, teachers, self), and to extract integrated scale scores. As a measurement model, Bauer’s model was designed for the purpose of generating integrated scores from item-level data across multiple informants, and it enabled researchers to evaluate processes that influence both the common and unique components of informant ratings.

The RTM applied in this study is mathematically equivalent to the models proposed in Bauer et al., (2013) and Rijmen et al., (2014). In this study, we use the logit link function for the dichotomous scores (e.g., pass vs. fail) based on the Rasch model. For polytomous responses, the model can be extended in a straightforward way by choosing the cumulative logit link (as in graded response modes) or the adjacent category logit link (as in partial credit models). As the simplest case, let $X_{pir}$ denote the binary score given by rater $r$, $r=1, ..., R$ on the $i$th item response, $i=1, ..., I$ for person $p$. Then the structure of the RTM is written as

$$\text{logit}(\text{Pr}(X_{pir} = 1 | \theta_p)) = \theta_p + \xi_{ir} + \beta_i - \rho_r$$

(1.4)
item and each rater. In particular, the RTM employed in this study considers parameterization of two types of rater effects: rater location \( \rho_r \) and rater variability \( \tau_{\zeta_r} \).

Figure 2-1 presents a representation of the RTM for the complete design when there are two raters \( (R=2) \) and three items \( (I=3) \). For example, in the figure, \( X_{p11} \) denotes the score given by Rater 1 on item 1 for person \( p \), and \( X_{p32} \) denotes the score given by Rater 2 on item 3 for person \( p \). The three types of latent variables, \( \theta_p \), \( \varepsilon_{ip} \), and \( \zeta_{rp} \), represent a common factor, \( I \) item-specific factors (one for each item), and \( R \) rater-specific factors (one for each rater) respectively. As shown in Figure 2-1, for example for a writing assessment, every score depends on the student’s overall writing ability (a common factor), a specific dimension for the writing task being asked (item-specific factor), and a specific dimension for the rater who gives scores (rater-specific factor).

There can be a couple of reasons for considering the RTM in the context of multiple rating situations: (1) we can incorporate a multiple ratings data structure explicitly by defining three factors for each response score (i.e., examinees, items, and raters), (2) we can evaluate the variability of the item-specific factors and rater-specific factors by relaxing the constant variances in these latent variables \( \varepsilon_{ip} \) and \( \zeta_{rp} \), and (3) we can avoid the rapid increase in the number of parameters for modeling the correlations between dimensions and we can aid conceptual interpretations of the latent variables. Regarding with the third reason, the orthogonality assumption in the RTM needs to be elaborated.
The outcome scores are assumed to be independent conditionally on three factors. As seen from Figure 2-1, the conditional independence is assumed among the raters, among the items, and particularly, between the raters and items, raters and examinees, and items and examinees. Note that each rating loads on one of each type of factor: the common factor, a specific factor for the rater, and a specific factor for the item. First, all ratings are allowed to load on the common factor $\theta_p$. This factor thus reflects the shared variability in the ratings across the raters and the items. It is considered to represent the consensus view of the examinee ability across raters and items. Next, the rater-specific factors, $\zeta_{1p}, \zeta_{2p}, \ldots, \zeta_{Rp}$, each affect only a single rater’s ratings and are assumed to be orthogonal to $\theta_p$ and to each other.

By imposing the constraint that the factors are orthogonal, we ensure that each rater factor captures variance that is unique to a specific rater and that is not shared with other raters. Given that the raters have passed the thorough training and qualification process, unique source of variation for each rater can be considered as the subjective biases. Finally, when there are multiple scores for the same response on a specific item, we anticipate that the ratings will be dependent not only due to the influence of the common factor but also due to the influence from that specific item. The item-specific factors, $\epsilon_{1p}, \epsilon_{2p}, \ldots, \epsilon_{Ip}$, account for this extra dependence. A given specific factor is defined to affect the responses of all raters to item $i$ but not to other items. These item-specific factors are also assumed to be orthogonal to one another and all other factors in the model. With these constraints, the item-specific factors capture covariation that is unique to a particular item. As noted by Marsh (1993), modeling item-specific factors for items rated by multiple raters is essential to avoid inflating the common factor variance. In addition, if the rater-item correlation is non-zero, the unique effects of the raters are not separated, complicating the assessment of specific sources of rater effects. Furthermore, items scored by disparate raters will correlate not only due to the common influence of the underlying true ability but also due to the fact that they are influenced by correlated raters.

In some cases, however, a researcher may have a theoretical rationale for permitting correlations among a subset of factors in the model. For instance, if the content of two items overlapped, such as in the testlets, then one may want to allow the item-specific factors for those items to be correlated. In effect, the correlation between these item-specific factors would account for the influence of a minor factor (e.g., same item stimuli) that jointly affects both items but that is narrower than the major factor of interest. For the RTM, failing to account for conditional dependence among the item-specific factors would be expected to distort the variance for the common factor, as this is the only other factor in the model that spans between raters. Introducing correlated specific factors may thus aid in avoiding model misspecifications that would otherwise adversely impact the estimation of the common factor. However, the trade-off is that the conceptual distinctions between the factors become blurred: the common factor no longer reflects all common variability across raters other than that unique to particular items, since some common variability is now accounted for by the correlations among item-specific factors. Similarly, if the ratings from the same scoring table are more similar to each other than ratings from other scoring tables, then the common factor would be required to account for the higher similarity of that specific table. Introducing a
correlation between the ratings among the raters from that specific table would account for the context of rating, enabling the common factor to integrate across-context ratings of raters more equitably. It is important to recognize, however, that introducing correlated rater-specific factors for raters originating from a common setting or context changes the definition of the common factor. The conceptual definition of the common factor is then affected by which rater-specific factors remain uncorrelated.

In summary, the assumption of orthogonality employed in the RTM enables us to define the rater-specific factors as the non-shared, unique components of variability in the rating scores of the raters. It also allows us to state that, within the model, the common and item-specific factors alone account for the shared variability across raters, with the common factor representing the broader construct of interest (e.g. writing ability) and the item-specific factors representing narrower contexts as defined by single item. Of course, if this is not a reasonable representation of the real-world contexts, one would have to investigate a more complex model.

**Variations**

As with all latent variable models, some constraints are necessary to set the scale of the latent variables. For identification, we first constrain the item difficulty of the last item as the negative sum of the rest of the item difficulties ($\beta_r = -\sum_{i=1}^{I-1} \beta_i$), and the rater severity of the last rater as the negative sum of the rest of the rater severities ($\rho_r = -\sum_{r=1}^{R-1} \rho_r$). Second, since in this study, all the item discrimination parameters are constrained to be 1 based on the Rasch model, the variances of the latent variables are freely estimated. These estimated variances can be viewed as the unique measurement error variances for each item and individual rater, which we call rater-specific or item-specific variance (Rabe-Hesketh & Skrondal, 2012, p. 608). Note that Equation (1.4) is the most general because it relaxes the constant variance assumptions for both crossed-dimensions, and all raters and items are allowed to have different measurement error variances: This will be called HEHE (heteroskedasticity for the item and raters).

More constrained versions of the RTM can be formulated by assuming homoskedastic variance either only for items, or only for raters, or both. First, with the homoskedastic variance only for rater-specific factors by retaining the previous model in the Equation (1.4), the assumptions are changed into $\varepsilon_{ip} \sim N(0, \tau_{\varepsilon}), \zeta_{rp} \sim N(0, \tau_{\zeta})$ and this is called HEHO (heteroskedasticity on item-specific factor and homoskedasticity on rater-specific factor). This assumption of constant rater-specific factor variance would imply that the decomposition of variance in the rating scores is identical across raters. Based on this HEHO model, we would estimate $I$ item locations as well as $I$ item-specific variance, $R$ rater locations, and two constant variances for general and rater-specific factors. Second, when heteroskedastic variance is allowed only for the rater-specific factor, the assumptions are changed into $\varepsilon_{ip} \sim N(0, \tau_{\varepsilon}), \zeta_{rp} \sim N(0, \tau_{\zeta})$ and this is called HOHE (homoskedasticity on item-specific
factor and heteroskedasticity on rater-specific factor). Following the HOHE model, we estimate \( I \) item locations, \( R \) rater locations, as well as \( R \) rater-specific variances, and two constant variances for the general and the item-specific factor. Last, when both the items and raters are assumed to have homoskedastic variance, the two crossed dimensions are assumed to follow \( \varepsilon_{ip} \sim N(0, \tau_{\varepsilon}) \), \( \zeta_{rp} \sim N(0, \tau_{\zeta}) \) and this is called HOHO (homoskedasticity on both item-specific factor and rater-specific factor). For the HOHO, it is assumed that all raters have the same measurement error variance \( \tau_{\zeta} \) and all items have the same measurement error variance \( \tau_{\varepsilon} \). In this model, we estimate \( I \) item locations, as well as \( R \) rater locations, and three constant variances for the general factor, the item-specific factor, and the rater-specific factor.

In terms of the rater effects, which are our main interest, rater severities (\( \rho_r \)) are estimated in all cases, which imply that there are systematic differences across raters in their levels of endorsement of the scores. With the HEHO and the HOHO, the rater-specific factor variances are assumed to be equal, and this implies that the decomposition of variance in the ratings is identical across raters. On the contrary, the HOHE and the HEHE allow us to investigate whether the raters have unique measurement error variances. As such, model comparison for the empirical data may provide substantively important information on whether and how the ratings of different raters differ.

In addition, to aid the practical interpretation of the heterogeneity among the raters, the median odds ratio proposed by Larsen and Merlo (2005) and Larsen, Petersen, Budtz-Jørgensen, & Endahl (2000) can be calculated. The median odds ratio is a useful measure of heterogeneity for random-intercepts models with normally distributed random intercepts. They consider repeatedly sampling two subjects with the same covariate values and forming the odds ratio comparing the subject who has the larger random intercept compared to the other subject. The median odds ratio is calculated as

\[
OR_{\text{median}} = \exp\{\sqrt{2\psi \Phi^{-1}(3/4)}\}
\]

where \( \psi \) is the estimated variance of the latent variable (i.e., \( \tau_{ei} \) for items and \( \tau_{\xi r} \) for raters) and \( \Phi^{-1} \) is the inverse of the normal distribution.

**Rater Assignment Design**

Aside from the assumptions of the RTM, we have to consider the rater assignment design as well as the item administration concerning the data. As the most ideal case, we can think of a situation where all the items are administered to all examinees and all raters provide ratings on all item responses. There is no designed missingness in the data under this complete (fully-crossed) design. The use of the complete design, however, is seldom feasible, because there are too many essays to be rated by each rater. Instead, in real world assessments, examinees are likely to respond to only a subset of items out of the pool of the items, and randomly assigned raters among the pool of raters will be able to give scores to those essays. The RTM is still applicable with incomplete design when the missingness comes into play.
However, for items rated by only one rater, the item-specific factors must be omitted from the model specification. For these items, the item-specific factor variance is conflated with random error. Similarly, raters who give scores only on single item must be discarded from the model specification.

In terms of the rater allocations, further modifications or restrictions on the model parameters can be imposed depending on the interchangeability of the raters. For example, in the GRE essay exams, examinees obtain two or three scores from two or three randomly chosen raters for each essay. If the third score is given by the automated scoring engine, that is structurally different from the other two scores given by the human raters, whereas the human raters are interchangeable. The RTM should be specified so that all parameters are constrained to be equally interchangeable across raters but should be allowed to differ across structurally different raters, such as an automated scoring engine from the human raters (Nussbeck, Eid, & Lischetzke, 2006).

**Relations to other models**

The RTM applied in this study is formally equivalent to the models proposed in Bauer et al., (2013) and Rijmen et al., (2014). In particular, the RTM HEHE in Equation (1.4) is equivalent with the unconditional model proposed in Bauer et al., (2013). By considering the fixed-effects (i.e., item locations and rater locations) as the means of the item-specific and rater-specific factors and by allowing different item discriminations holding the standard normal distributions for the latent variables, the mathematical relationship between the two models can be easily established. However, the RTM applied in this study has conceptual and practical differences from Bauer’s model. In Bauer’s model, the primary interest lies in abstracting the common element across multiple informants’ ratings while isolating the unique perspectives and potential biases of the individual reporters (e.g., self, teachers, parents). In the contexts of their interest, the unique perspective of each informant is natural depending on the various informants’ contexts and relationships with examinees. However, the multiple ratings data of our interest deals with more homogeneous groups of trained raters who have likely passed a thorough training and qualification process. Thus, we consider the differences from different raters as subjective biases of the raters, which need to be minimized, rather than recognizing them as the unique perspective of each rater. Another complexity of our context has to do with the missingness in the data. In the real world, the designs used in large-scale assessments are necessarily incomplete due to the rater allocation and the item assignment. In sum, with the application of the RTM, we would be more interested in investigating the effect of ignoring differences among the raters, particularly, their rater-specific measurement variances, and the effect of incompleteness (i.e. missingness rate of the data). In this study, these issues are examined via two sets of simulation studies.

Another interesting relationship with the RTM can be found from the G-theory perspective (Brennan, 1992; Briggs & Wilson, 2007). Note that the RTM HEHE described above implies that the variance of the latent variables is additively decomposed into multiple components involving $\varepsilon_p$, $\zeta_p$, and $\theta_p$. This is analogous to a G-theory approach that treats all components as random-effects. G-theory focuses on the magnitude of sampling variability.
due to items, raters, and so forth, and their combinations, providing estimates of the magnitude of measurement error in the form of variance components. We can compute the proportion of the variance determined by item, rater, and general factors based on the RTM HEHE as follows (Jeon, Rabe-Hesketh, & Rijmen, 2013b; Nussbeck, Eid, & Lischetzke, 2006; Shavelson, Baxter, & Gao, 1993):

\[
\gamma_{pr}^R = \frac{\text{Var}(\zeta_{rp})}{\text{Var}(\zeta_{rp}) + \text{Var}(\epsilon_{ip}) + \text{Var}(\theta_p)},
\]

\[
\gamma_{pr}^I = \frac{\text{Var}(\epsilon_{ip})}{\text{Var}(\zeta_{rp}) + \text{Var}(\epsilon_{ip}) + \text{Var}(\theta_p)},
\]

\[
\gamma_{pr}^G = \frac{\text{Var}(\theta_p)}{\text{Var}(\zeta_{rp}) + \text{Var}(\epsilon_{ip}) + \text{Var}(\theta_p)}.
\]

Following Equation (1.6), coefficients \(\gamma_{pr}^R\) and \(\gamma_{pr}^I\), the proportion of true variance determined by the rater-specific and item-specific factors, can be identified as so-called rater-specificity and item-specificity coefficients, respectively. Sampling variability due to raters speaks to a traditional concern about the inter-rater reliability. Sampling variability due to items traditionally has been thought of as related to internal consistency reliability. These coefficients also correspond to the method-specificity and consistency coefficients used in the multitrait-multimethod literature (Campbell & Fiske, 1959), which provides convergent and discrimination validity evidence for latent trait and method effects.

### 2.2.3. Estimation

Maximum likelihood (ML) estimation of multidimensional models is computationally demanding due to the high-dimensional integrals over latent variables in the likelihood functions. Although the integration dimensionality can be reduced by relying on the conditional independence relations between the primary and specific dimensions in the bifactor model (Rijmen, 2009), the computational cost of RTM remains high due to the crossed structure of the specific dimensions. ML and Bayesian MCMC estimation approaches can accommodate missing data. Both ML and MCMC estimation include approaches for partial data under the assumption that the missing data are missing at random. For estimating the proposed RTM, a Bayesian approach using Markov chain Monte Carlo (MCMC) was implemented in WinBUGS 1.4.3 (Lunn, Thomas, Best, & Spiegelhalter, 2000). Bayesian estimation by MCMC approximates maximum likelihood when priors are selected to be noninformative (Edwards, 2010).

In order to implement MCMC algorithm using WinBUGS, prior distributions must be specified for all parameters, which include the item difficulties and rater severities as well as the variance of the person proficiency distribution, item-specific variances, and rater-specific variances. Although each parameter could potentially have many different prior distributions, this study limits its scope to the simple and straightforward commonly used ones such as the
conjugate priors that ensure the posterior distribution belong to the same family as the prior. More specifically, assuming a normal distribution is a standard practice for the fixed-effects parameters, and the conjugate prior for the variance of the normal distribution is the inverse-gamma distribution. Following mildly informative prior distributions were specified to estimate the proposed model in this study,

\[
\begin{align*}
\beta_i &\sim N(0,1), \quad i = 1,\ldots,I, \\
\rho_r &\sim N(0,1), \quad r = 1,\ldots,R, \\
\theta_p &\sim N(0,\psi), \\
\psi &\sim Inverse-Gamma(1,1), \\
\varepsilon_{ip} &\sim N(0,\tau_{\varepsilon_i}), \\
\tau_{\varepsilon_i} &\sim Inverse-Gamma(1,1), \\
\zeta_{rp} &\sim N(0,\tau_{\zeta_r}), \\
\tau_{\zeta_r} &\sim Inverse-Gamma(1,1).
\end{align*}
\]

2.3. Simulation Study

2.3.1. Data Generation

Fifty complete (fully-crossed) data sets were generated using the fixed values based on the RTM HEHE in Equation (1.4). The complete (fully crossed) design, where each response is scored by every rater without any missingness, is taken as an ideal example. In the data generation, the R software (R Development Core Team, 2013) was used. The number of examinees was deliberately set as 1,080 for the simulation design following DeCarlo (2008, 2010), which will be explained later, and the latent abilities of the examinees were generated from a normal distribution with zero mean and variance of 2. The number of items was 10, and item locations were set between -2.9 and 3.1 and fixed across the data sets. Similarly, the number of raters was 10, and the rater severities were fixed across the data sets, which were set between -2.9 and 3.1 as well. The mean of the item locations and the rater locations were constrained to be zero. Heterogeneous variances for the different raters and items were set between 0.3 and 3.0 to be equally distributed and fixed across the simulations. Two sets of simulation studies were carried out to evaluate the parameter recovery of the RTM under the different conditions. In order to evaluate the feasibility of the RTM, the first simulation study condition varied the missingness rate due to rater allocation and item administration (M=0%, 75%, and 96%), and the second simulation study made incorrect assumptions on the measurement error variance of the two crossed dimensions (i.e., generated by HEHE and analyzed with HOHO, HOHE, and HEHO).

Complete (fully crossed) designs are not used in practice for large-scale assessments due to the large number of essays that would need to be scored, hence, the use of incomplete designs is more popular. For example, in the GRE scoring procedure, examinees respond on
two writing tasks (i.e., “Analyze an Issue” and “Analyze an Argument”) out of a pool of items and each essay is typically scored by two raters out of a pool of raters (Educational Testing Service, 2011). In order to vary the missingness, an unbalanced incomplete design first employed by DeCarlo (2010) was modified and used in this study. At the first stage, data were created from the fully crossed design and missing values were inserted according to the unbalanced incomplete design. Unlike DeCarlo (2008, 2010), who limited the context where examinees respond to a single writing task thus reducing the item facet, we consider multiple tasks, making the item assignment an important element. In real world assessments, it is typical that raters are trained to give scores for some specific item types, and all possible pairings of the raters are hard to implement particularly in the context of multiple tasks.

For the unbalanced incomplete design employed in this study, blocks of all possible rater pairs were created and all other data points were converted to missing. Consider the GRE example in this study that 1,080 examinees responded to 2 out of 10 writing tasks, where each essay is given score by 2 raters among the pool of 10 raters. Note that data that are missing by design are missing completely at random (Rubin, 1976) and the necessary conditions below were studied in Fleiss (2011):

\[
\begin{align*}
gr &= nk, \\
g &\leq n, \\
\lambda (g - 1) &= r(k - 1),
\end{align*}
\]

where \(g\) is the number of raters, \(r\) is the number of essays scored by each rater, \(n\) is the number of essays, \(k\) is the number of raters that score each essay (i.e., the block size), and \(\lambda\) is the number of essays scored by each pair of raters. For the design examined here, it follows from the first condition that \(10\times r = (1080\times 2)\times 2\), resulting in \(r = 432\). Further, the third condition suggests \(\lambda \times 9 = 432\), resulting in \(\lambda = 48\). Note that the number of examinees, 1,080, was deliberately chosen to make \(\lambda\) a whole number. In summary, the incomplete aspect of the design is that (a) each essay is scored by only 2 out of 10 raters, and (b) each rater is paired with specific other raters, whereas the balanced aspect is that (a) each essay is scored by 2 raters, and (b) each rater scores 432 essays.

Following the GRE essay example, the unbalanced incomplete design deliberately converted the complete (fully crossed) design to have an amount of missingness similar to typical examples (see Table 2-2). During the allocation of the items, a spiral-like arrangement, with the nine pairs (1,2), (2,3), (3,4),…, (9,10), was used. First, examinees were divided into 10 different groups, each having 108 examinees, and each group of examinees was assumed to respond on two sequential items and get scores from a specific rater pair. When using fewer than all of the possible pairs of raters, one has to take care that the raters are all “connected”, so that, the parameters are on a common scale and can be compared across all of the raters.

Table 2-2 depicts the allocation of items and raters under the GRE design with the number of data points in each cell. As illustrated in Table 2-2, for example, 108 examinees in the exam group 1 responded Item 1 and Item 2, and got scores from a rater pair, Rater 1 and 10. From rater’s perspective, Rater 1 gives scores on exam group 1 and 10, who responded Item 1, 2, and 10, and give scores to total 432 essays. It was assumed that each rater has been
trained to give scores for three items. There were ten unique rater pairs instead of all possible 45 rater pairs, one for each examinee group, and the number of ratings from each unique rater pair was same as 432, which was 216 for each item. Although all possible rater pairs could not be used for the simulation data due to the missingness, we tried to make rater pairs deliberately connected in some ways. For example, under the GRE design, Rater 1 and Rater 4 are connected by having Rater 10 as the common rater pair, and Rater 8 and Rater 3 are connected by having Rater 7 as the common rater pair. The connectedness between the raters is illustrated in Figure 2-2. This design was used because it is simple to implement, and allows one to create unbalanced incomplete data with some control over how many essays each rater scores. In summary, three sets of simulation data which vary by missingness rate were generated and analyzed:

1) The Complete design (Missingness rate = 0%): Examinees respond on every item, and every essay is scored by every rater, which results in 100% valid scores (1,080*10*10=108,000 data points)

2) The Pseudo-GRE design (Missingness rate = 75%): As a bridge situation between the complete design and the GRE design, examinees respond five items out of ten items, and five raters out of ten raters based on the spiral design give scores independently for each essay, which results in 25% valid scores (1,080*5*5=27,000 data points)

3) The GRE design (Missingness rate = 96%): Examinees respond to only two items out of ten, and two raters out of ten raters based on the spiral design, give scores independently for each essay, which results in only 4% valid scores (1,080*2*2=4,320 data points)

<table>
<thead>
<tr>
<th>Table 2-2. Allocation of items and raters in the GRE design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam Group</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>1 (item 1,2)</td>
</tr>
<tr>
<td>2 (item 2,3)</td>
</tr>
<tr>
<td>3 (item 3,4)</td>
</tr>
<tr>
<td>4 (item 4,5)</td>
</tr>
<tr>
<td>5 (item 5,6)</td>
</tr>
<tr>
<td>6 (item 6,7)</td>
</tr>
<tr>
<td>7 (item 7,8)</td>
</tr>
<tr>
<td>8 (item 8,9)</td>
</tr>
<tr>
<td>9 (item 9,10)</td>
</tr>
<tr>
<td>10 (item 10,1)</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
As a descriptive statistic, the agreement rates between the rater pairs were calculated for each simulation design. For the Complete design, for example, the agreement rates for all possible 45 rater pairs were calculated, and the agreement rate between the two raters ranged from 37.79% to 77.71%, with a median of approximately 67%. Using the generating values, the highest agreement rate came from the Rater 3 and Rater 8 pair, and the lowest agreement rate came from the Rater 2 and Rater 3 pair across all the replicates. When the missingness came into play, this pattern became different due to the connectedness between the raters. For the Pseudo-GRE design, there were 40 unique rater pairs who gave results (i.e., rater pair-specific missing rate > 0.00), while the scores from the other 5 pairs (Rater 1 and Rater 6, Rater 2 and Rater 7, Rater 3 and Rater 8, Rater 4 and Rater 9, and Rater 5 and Rater 10) turned into missing. Under this design, the agreement rate between the two raters turned to be lower than the Complete design, ranging from 10.32% to 71.99%, with a median of approximately 38%. For the GRE design, as designed in Table 2-2, only 10 rater pairs gave the results, while the scores from other 35 rater pairs became missing. The agreement rates under the GRE design showed larger range than the Complete design, ranging from 23.15% to 85.19%, with a median around 63%. Therefore, it is expected that the missingness brought by the design would considerably impact the parameter recovery results.

### 2.3.2. Analysis

Once the data sets were generated, the true model, the RTM HEHE was fit using MCMC estimation across 50 Complete design replicates. In order to examine the impact of the missingness, the same RTM HEHE was analyzed across 50 simulated GRE design replicates and Pseudo-GRE design replicates. For each case, bias and root mean square error (RMSE) were computed to assess the parameter recovery. Furthermore, in order to examine the impacts of ignoring the heterogeneity in the variances of the item-specific and rater-specific factors, three misspecified models that hold the homoskedasticity assumption either on item or rater or both (HOHO, HOHE, and HEHO) were applied to 50 Complete design replicates. Throughout all the analyses, WinBUGS was run using three chains with 7,000 iterations after discarding 3,000 burn-in iterations. In order to check convergence, time-series plots were monitored and three chains with differing initial values were specified. Convergence of the three chains was examined using the \( \hat{R} \) index proposed by Gelman and Rubin (1992) with a critical value of 1.1.
2.3.3. Results

Simulation Study I: Impact of Missingness Rate on Parameter Recovery

Across 50 replicates for three different missingness rate conditions, the true model, the RTM HEHE was fitted. The estimated bias and RMSE are given in Table 2-3 for each model parameter by the missingness rates. There were 39 parameter estimates in total, including 18 fixed-effects (nine item difficulties, $\beta_i$, $i=1,\ldots,9$, and nine rater severities, $\rho_r$, $r=1,\ldots,9$), and 21 random-effects (10 item-specific measurement variances, $\tau_{\epsilon i}$, $i=1,\ldots,10$, 10 rater-specific rater measurement variances, $\tau_{\epsilon r}$, $r=1,\ldots,10$, and a variance of the person proficiency distribution, $\psi$).

Table 2-3. Bias and RMSE of the Simulation Study by Missingness Rate Design

<table>
<thead>
<tr>
<th></th>
<th>Complete (0% missingness)</th>
<th>Pseudo-GRE (75% missingness)</th>
<th>GRE 96% missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.024</td>
<td>0.039</td>
<td>-0.006</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.011</td>
<td>0.029</td>
<td>0.054</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.054</td>
<td>0.061</td>
<td>0.022</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.070</td>
<td>0.075</td>
<td>-0.099</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.026</td>
<td>0.043</td>
<td>-0.005</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.079</td>
<td>0.089</td>
<td>0.114</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-0.050</td>
<td>0.060</td>
<td>-0.050</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0.032</td>
<td>0.048</td>
<td>0.063</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-0.028</td>
<td>0.038</td>
<td>-0.064</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>-0.012</td>
<td>0.035</td>
<td>-0.036</td>
</tr>
<tr>
<td>$\rho_2$</td>
<td>-0.013</td>
<td>0.045</td>
<td>-0.035</td>
</tr>
<tr>
<td>$\rho_3$</td>
<td>0.012</td>
<td>0.039</td>
<td>-0.080</td>
</tr>
<tr>
<td>$\rho_4$</td>
<td>0.046</td>
<td>0.055</td>
<td>0.070</td>
</tr>
<tr>
<td>$\rho_5$</td>
<td>0.045</td>
<td>0.052</td>
<td>0.069</td>
</tr>
<tr>
<td>$\rho_6$</td>
<td>-0.004</td>
<td>0.031</td>
<td>0.110</td>
</tr>
<tr>
<td>$\rho_7$</td>
<td>0.017</td>
<td>0.032</td>
<td>0.021</td>
</tr>
<tr>
<td>$\rho_8$</td>
<td>-0.027</td>
<td>0.050</td>
<td>-0.062</td>
</tr>
<tr>
<td>$\rho_9$</td>
<td>-0.065</td>
<td>0.072</td>
<td>-0.058</td>
</tr>
<tr>
<td>$\tau_{\epsilon 1}$</td>
<td>0.009</td>
<td>0.074</td>
<td>0.043</td>
</tr>
<tr>
<td>$\tau_{\epsilon 2}$</td>
<td>0.020</td>
<td>0.114</td>
<td>0.048</td>
</tr>
<tr>
<td>$\tau_{\epsilon 3}$</td>
<td>0.059</td>
<td>0.102</td>
<td>0.018</td>
</tr>
<tr>
<td>$\tau_{\epsilon 4}$</td>
<td>0.056</td>
<td>0.112</td>
<td>0.114</td>
</tr>
<tr>
<td>$\tau_{\epsilon 5}$</td>
<td>-0.081</td>
<td>0.181</td>
<td>-0.317</td>
</tr>
<tr>
<td>$\tau_{\epsilon 6}$</td>
<td>0.165</td>
<td>0.239</td>
<td>-0.035</td>
</tr>
<tr>
<td>$\tau_{\epsilon 7}$</td>
<td>-0.062</td>
<td>0.173</td>
<td>-0.083</td>
</tr>
<tr>
<td>$\tau_{\epsilon 8}$</td>
<td>-0.019</td>
<td>0.133</td>
<td>0.032</td>
</tr>
</tbody>
</table>
Overall, under the Complete design where the missingness rate was zero, the bias values were smaller compared with the corresponding values from the Pseudo-GRE design or the GRE design. By missingness rate, it was clear that both bias and RMSE values got bigger as the missingness rate increased. Across different missingness rates, both positive and negative bias were observed, which suggests that there is no clear pattern of overestimation or underestimation. The comparison of the bias and RMSE values for the three different missingness rates for the fixed-effects and the random-effects parameters is also illustrated in Figure 2-3. As shown, the fixed-effects parameters tended to have smaller bias and RMSE values compared with the random-effects parameters across all conditions. In particular, under the Complete design, the fixed-effects parameters were recovered well, with the median of the absolute value of the bias equal to 0.028 and the median of the RMSE equal to 0.044. In contrast, the random-effects parameters were not recovered so well, with the median of the absolute value of the bias equal to 0.040 and the median of the RMSE equal to 0.114. This pattern was consistent across the different missingness rates. For the Pseudo GRE design with 75% missingness rate, the median absolute bias was 0.056 and the median RMSE was 0.089 for the fixed-effects and corresponding values were 0.076 and 0.300 for the random-effects. Under the GRE design with 96% missingness, median absolute bias was 0.161 and the median RMSE was 0.339 for the fixed-effects and the corresponding values were 0.553 and 1.486 for the random-effects.

In summary, parameter recovery appears greatly sensitive to the missingness rate. This can be interpreted as being due to the weak connectedness between the rater pairs and the item allocations, which is in line with DeCarlo (2010), although the contexts and analysis models are different. In his study, the results suggested that estimation is affected by a lack of balance and by having less than full linkage. In this study, given that the missingness resulted in a weak linkage as described in the different patterns of the agreement rates between the rater pairs, allocation of the items and rater assignments, particularly, how the raters are connected to each other, are crucial factors to recover the true values of the RTM HEHE.
Figure 2-3. Bias and RMSE of the Simulation Study by Missing Rate Design
Simulation Study II: Impact of Homoskedasticity Assumption on Parameter Recovery

Another primary goal of the simulation study was to investigate the effects of ignoring differences among the raters, particularly, their rater-specific measurement variances. When the participating raters have different measurement error variances, assuming the homogeneous variances across all the raters may lead to inaccurate description of not only the rater effects, but also the item properties and the person proficiencies. Thus, across the 50 Complete design replicates, three types of misspecified assumptions, HOHO, HOHE, HEHO were analyzed and compared with the correct assumption, HEHE.

The estimated DIC values were consistently lower for the generating model than three models with incorrect assumptions. This result suggests that the estimation is doing a good job of identifying the correct model. Specifically, the DIC values of HEHE was smaller than ones of HOHO, HOHE, and HEHO by more than 5 units, which has been considered the minimum cut-off representing a substantial drop to support the better fit (Li, Bolt, & Fu, 2006). The median DIC value of HEHE was 81539.9, and the ones from HOHO, HOHE, and HEHO were 82319.85, 81940.6, and 81944.0, respectively. In particular, constraining both rater-specific variances and item-specific variances as constant (HOHO) always showed the worst fit compared to the other models.

As summarized in Table 2-4 and illustrated in Figure 2-4, there were not large differences in bias and RMSE values for the fixed-effects parameters between the misspecified models and the generating model. However, for the random-effects parameters, incorrect models resulted in larger bias and RMSE values. In particular, the variance of the person ability distribution ($\psi$) tended to be slightly underestimated when the misspecified models were fit. In general, the magnitude of bias appeared similar to the correct model specification (HEHE) with missingness rate 75%.

Table 2-4. Bias and RMSE of the Simulation Study by Model Misspecification

<table>
<thead>
<tr>
<th></th>
<th>HOHO</th>
<th></th>
<th>HOHE</th>
<th></th>
<th>HEHO</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.035</td>
<td>0.047</td>
<td>0.031</td>
<td>0.044</td>
<td>-0.017</td>
<td>0.035</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.012</td>
<td>0.030</td>
<td>0.014</td>
<td>0.031</td>
<td>0.008</td>
<td>0.030</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.016</td>
<td>0.034</td>
<td>0.018</td>
<td>0.035</td>
<td>0.048</td>
<td>0.056</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.099</td>
<td>0.102</td>
<td>-0.101</td>
<td>0.104</td>
<td>-0.072</td>
<td>0.077</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.034</td>
<td>0.046</td>
<td>-0.027</td>
<td>0.041</td>
<td>0.016</td>
<td>0.036</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.221</td>
<td>0.224</td>
<td>0.214</td>
<td>0.217</td>
<td>0.102</td>
<td>0.110</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-0.020</td>
<td>0.034</td>
<td>-0.019</td>
<td>0.033</td>
<td>-0.048</td>
<td>0.058</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0.048</td>
<td>0.058</td>
<td>0.048</td>
<td>0.058</td>
<td>0.032</td>
<td>0.047</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-0.055</td>
<td>0.061</td>
<td>-0.054</td>
<td>0.060</td>
<td>-0.035</td>
<td>0.043</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>0.046</td>
<td>0.055</td>
<td>-0.126</td>
<td>0.129</td>
<td>-0.034</td>
<td>0.044</td>
</tr>
<tr>
<td>$\rho_2$</td>
<td>0.120</td>
<td>0.128</td>
<td>-0.001</td>
<td>0.029</td>
<td>0.044</td>
<td>0.053</td>
</tr>
<tr>
<td>$\rho_3$</td>
<td>-0.067</td>
<td>0.076</td>
<td>-0.030</td>
<td>0.056</td>
<td>0.129</td>
<td>0.137</td>
</tr>
</tbody>
</table>
In order to illustrate the impact of misspecified homoskedasticity assumption in more
detail, boxplots of the median odds ratios were generated in comparison with the true median
odds ratio. In Figure 2-5 and Figure 2-6, the distribution of the median odds ratios by items
and raters are shown respectively with the white dot for the true median odds ratio. In general,
the correct model (HEHE) recovered the true median odds ratio well for both items and raters,
except for the ones with the highest extant such as Item 6 and Rater 2. Interestingly, the
median odds ratios were recovered well when the heterogeneous variance of the items and
raters was not modeled. For example, under the HEHO where the constant variance was
imposed for the raters, the median odds ratios by items were recovered with similar success as
the correct model (HEHE). This pattern was consistent when the constant variance was
imposed for the items (HOHE) and median odds ratios by raters were calculated. Thus, in
terms of the recovery of the median odds ratio by the items or raters, it looked pretty robust to
the misspecified homoskedasitic variances of another crossed-dimensions.
Figure 2-4. Bias and RMSE of the Simulation Study by Misspecification
Figure 2-5. Distribution of the Median Odds Ratio by Items

Figure 2-6. Distribution of the Median Odds Ratio by Raters
2.4. Empirical Example

2.4.1. Data Source

To illustrate the use of the RTM in real world assessments, multiple ratings data from the Carbon Cycle project (Jin & Anderson, 2008), which were collected during the 2008–2009 school year, were used. The empirical data utilized an unbalanced incomplete design in terms of the assignment of items to students and allocation of the item responses to raters, as could be expected in practical multiple rating situations. Compared to the unbalanced incomplete design employed in the simulation study, the empirical data set added more incomplete aspects because the number of the responses each rater gives scores is different. As Wilson and Hoskens (2001) noted, if a rater only rates a single item, and no other rater rates that item, then the rater severity and the item difficulty will be confounded. Thus, we only included ratings by a certain rater pair if they rated more than 5 items and when an individual item had more than 20 ratings. Therefore, the resulting item response data contained double ratings of the 10 written assessment items that were answered by a total of 205 students from the states of Michigan and Washington. The original data was scored polytomously according to the four levels of achievement defined in the theoretical learning progression framework. To make it easier to illustrate the use of the RTM used in this study, we dichotomized the item response data: for example, scores 1 and 2 were recoded into 0, and scores 3 and 4 were recoded into 1.

There were 8 raters in total, and the data used for the analysis included two ratings from two different raters on the same response. The allocation of the 8 raters to the student responses was designed to overlap some of the items in the assessment. The pairs of raters in the data set are shown in Table 2-5 with the entries in the matrix showing the number of ratings on which the raters were paired. Note that the pattern of data incidence is quite asymmetric unlike the simulation data sets. There were 15 unique rater pairs, and the number of ratings from each unique rater pair ranged from 3 to 172. The agreement rate between raters ranged from 28.57% to 100%, with a median of 84.76%. It is noticeable that the agreement rate between each rater pair was high and the distribution of agreement rates by item was relatively even.

<table>
<thead>
<tr>
<th>Rater</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>172</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>71</td>
<td>21</td>
<td>3</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>70</td>
<td>0</td>
<td>105</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>41</td>
<td>0</td>
<td>63</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>
2.4.2. Analysis

In this study, MCMC estimation was implemented in WinBUGS. WinBUGS was run using three chains with 7,000 iterations with a burn-in of 3,000 iterations. In order to check convergence, time-series plots were monitored and three chains with different initial values were specified. Convergence of the three chains was examined using the \( \hat{R} \) index proposed by Gelman and Rubin (1992) with a critical value of 1.01.

For model selection, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) were used. The AIC and BIC indices are reported, which are defined as:

\[
\begin{align*}
AIC &= D(\xi) + 2m, \\
BIC &= D(\xi) + m(\log N),
\end{align*}
\]

where \( D(\xi) \) is the posterior mean of the deviance, \( \xi \) represents all parameters under the model, \( m \) refers to the number of estimated parameters, and \( N \) indicates the sample size.

2.4.3. Results

Table 2-6 presents a comparison of the model fit for the empirical Carbon Cycle project data and the number of parameters and dimensions. For this empirical example, the RTM HOHO was the best fitting model, which suggests that relaxation of the homoskedasticity is not required for both items and raters. When compared with the MFRM (Linacre, 1994) in Equation (1.1) that ignored the conditional dependence between multiple ratings by excluding random variables associated with items and raters, all of the RTMs fitted better in terms of AIC and BIC. This suggests that conditional dependence from items and raters is not negligible and should be taken care of. Further, comparison between the RTM HOHE and the RTM HEHO reveals that relaxing homogeneity for items (HEHO) is more crucial than relaxing homogeneity for raters (HOHE).

<table>
<thead>
<tr>
<th>RTM</th>
<th>#Par</th>
<th>#Dim</th>
<th>Deviance</th>
<th>AIC</th>
<th>BIC</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOHO</td>
<td>21</td>
<td>3</td>
<td>794.57</td>
<td>830.12</td>
<td>896.58</td>
<td>1076.12</td>
</tr>
<tr>
<td>HOHE</td>
<td>28</td>
<td>10</td>
<td>873.45</td>
<td>918.23</td>
<td>1007.95</td>
<td>1138.70</td>
</tr>
<tr>
<td>HEHO</td>
<td>30</td>
<td>12</td>
<td>819.99</td>
<td>872.63</td>
<td>968.99</td>
<td>1093.95</td>
</tr>
<tr>
<td>HEHE</td>
<td>37</td>
<td>19</td>
<td>804.25</td>
<td>870.07</td>
<td>989.69</td>
<td>1088.52</td>
</tr>
<tr>
<td>Baseline</td>
<td>MFRM</td>
<td>19</td>
<td>1</td>
<td>963.47</td>
<td>1001.47</td>
<td>1064.61</td>
</tr>
</tbody>
</table>

Table 2-7 lists the MCMC parameter estimates of the four types of the RTM analyses. First, the fixed-effects item difficulties appeared to be similar across four different
assumptions. Item 1 was the easiest item and Item 6 was the most difficult item regardless of the models. In contrast, the fixed-effects rater severities yielded only a small difference between when the homoskedasticity assumption holds (HOHO, HEHO) and when the homoskedasticity assumption gets relaxed (HOHE, HEHE). Although the ranking of the rater severities looks fairly consistent, a significance test of the rater severities resulted in different conclusions. For example, under the homoskedasticity assumptions (HOHO and HEHO), Rater 8 showed statistically significant severity (i.e. different from zero). When the homoskedasticity assumption is relaxed, however, this rater’s severity was no longer statistically significantly different from zero.

Second, regarding the random effects, the estimated variance of the general factor (i.e., person proficiency, \( \psi \)) was similar across models, except for HOHE. Given that HOHE showed the worst model fit for this empirical data, it seems inappropriate to consider different measurement variances for each rater. Based on the best-fitting HOHO model, the estimated variances for the general factor, item-specific factor, and rater-specific factor decrease in that order. From the G-theory perspective, based on the best-fitting HOHO model, the proportion of variance is decomposed into the general factor, which is the largest at 0.67, then the item-specific factor at 0.25, and the rater-specific factor, which is the smallest at 0.07. Given that the item-specificity and rater-specificity coefficients are low compared with the general factor, we can say that the assessment used in the Carbon Cycle project was valid to assess the environmental literacy as intended. It also implies that multiple ratings from the Carbon Cycle project vary to a larger degree between the items than between the raters. Specifically, the low proportion for the rater-specific factor was due to the homogeneity of the raters, who were trained with the same scoring rubric and who passed the reliability check process. On the other hand, the examinee’s scores were less consistent across the sample of different items; some items were easy to do for some examinees but not for other students.

Table 2-7. Parameter estimates and standard errors for the Carbon Cycle project data

<table>
<thead>
<tr>
<th></th>
<th>HOHO</th>
<th>HOHE</th>
<th>HEHO</th>
<th>HEHE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>Est.</td>
<td>SE</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>-1.53</td>
<td>0.40</td>
<td>-1.43</td>
<td>0.35</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>0.02</td>
<td>0.42</td>
<td>-0.02</td>
<td>0.37</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>-0.32</td>
<td>0.30</td>
<td>-0.30</td>
<td>0.26</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>0.06</td>
<td>0.39</td>
<td>0.07</td>
<td>0.33</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>0.34</td>
<td>0.45</td>
<td>0.32</td>
<td>0.39</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>1.54</td>
<td>0.32</td>
<td>1.42</td>
<td>0.27</td>
</tr>
<tr>
<td>( \beta_7 )</td>
<td>-0.78</td>
<td>0.32</td>
<td>-0.71</td>
<td>0.27</td>
</tr>
<tr>
<td>( \beta_8 )</td>
<td>-0.21</td>
<td>0.38</td>
<td>-0.12</td>
<td>0.33</td>
</tr>
<tr>
<td>( \beta_9 )</td>
<td>0.54</td>
<td>0.29</td>
<td>0.47</td>
<td>0.26</td>
</tr>
<tr>
<td>( \beta_{10} )</td>
<td>0.35</td>
<td>0.33</td>
<td>0.30</td>
<td>0.28</td>
</tr>
<tr>
<td>( \rho_1 )</td>
<td>0.21</td>
<td>0.25</td>
<td>0.46</td>
<td>0.23</td>
</tr>
<tr>
<td>( \rho_2 )</td>
<td>-0.24</td>
<td>0.29</td>
<td>-0.06</td>
<td>0.28</td>
</tr>
<tr>
<td>( \rho_3 )</td>
<td>0.62</td>
<td>0.32</td>
<td>0.65</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Overall, from the best-fitting HOHO results, Rater 5 was the most lenient rater, and Raters 4 and 8 were the most severe raters. Although the model fit result suggests that raters are similar in terms of the rater variability, the median odds ratio given in Equation (1.5) was calculated for comparison with the second-best fitting HEHE model estimates. From two randomly chosen examinees of the same rater for the given item, the median odds ratio between the examinee with the highest propensity to get a satisfactory (1) score and the examinee with the lowest propensity to get a unsatisfactory (0) score is estimated to be 2.22, 2.24, 2.56, 2.07, 2.92, 2.39, 3.56, and 4.32 respectively for 8 raters. Again, although Rater 7 and 8 show relatively large median odds ratio, comparison of the model fit suggests that the heterogeneity among the raters is not significant. For comparison, when the two randomly chosen examinees from the same item for the given rater, one finds 3.23, 3.28, 3.41, 3.75, 2.83, 3.66, 4.15, 4.43, 2.35, and 3.48 for the items. These values are larger than the median odds ratio from the same rater. This is in line with the model fit results that HEHO is favored over HOHE.
2.5. Conclusion and Discussion

Under multiple rating situations in which different raters give independent scores for the same responses (e.g., essay), the conditional independence assumption in the conventional IRT models is violated. Within the context of multiple ratings, the main concern is how to deal with the conditional dependencies among the ratings in an appropriate way. Bifactor item response models have mostly been developed to accommodate conditional dependence in different situations, such as testlets or item bundles. In this study, we have introduced the use of the RTM that extends bifactor model to incorporate a cross-classified structure (e.g., items and raters) in addition to the general dimension for the multiple ratings data. Specifically, four types of the RTMs depending on relaxation of the homoskedasticity assumption on a cross-classified structure (HOHO, HOHE, HEHO, HEHE) have been described and investigated.

The designs used in real world large-scale multiple rating situations are necessarily incomplete due to the practical situation of rater and item assignment. Thus, investigating the effect of incompleteness was examined in the first simulation study by varying the missingness rate in the data (i.e. the Complete design with 0% missingness, the Pseudo-GRE design with 75% missingness, and the GRE design with 96% missingness). In general, both bias and RMSE values became larger as the missingness rate increased. In terms of bias, both positive and negative bias were observed, which suggests that there is no clear pattern of overestimation or underestimation. In particular, our GRE design produced large biases and RMSEs due to the weak connectedness between the rater pairs and item allocations. Given that the missingness resulted in a weak linkage as described in the different patterns of the agreement rates among the rater pairs, allocation of the items and rater assignments (particularly, how the raters are connected to each other to have a common metric for the comparison of them) are crucial factors that affected the recovery of the generating values of the RTM HEHE under the larger missingness condition. This result, for a common rater design like the GRE design, deserves more study.

Another primary goal of the simulation study was to investigate the effect of ignoring differences among the raters, particularly, their rater-specific measurement variances. When the participating raters have different measurement error variances, assuming the homogeneous variances across all the raters may lead to inaccurate description of not only the rater effects, but also the item properties and the person proficiencies. Simulation study revealed that imposing homogeneous variances on either items or raters or both (HOHO, HOHE, HEHO) resulted in similar magnitudes of bias and RMSE values compared with the true HEHE for the fixed-effects parameters. For the random-effects parameters, incorrect models resulted in larger bias and RMSE values. In particular, the variance of the person ability distribution (\( \psi \)) tended to be slightly underestimated when the misspecified models were fit. In general, the magnitude of bias appeared similar to the correct model specification (HEHE) with missingness rate 75%.

In addition, for the empirical example of the Carbon Cycle project data, this study found that the RTM HOHO fitted better than any other model, which implies that
participating raters were not significantly different in their measurement variances although they were different in their severities.

Finally, this study was a preliminary investigation of the use of the RTM for multiple ratings data. Although the use of the RTM has been investigated via two simulation studies and empirical data analysis, it would be interesting to apply the RTMs to other different multiple rating situations. A major limitation is that the RTMs employed in this study only dealt with the dichotomous scores (e.g., satisfactory vs. unsatisfactory), which is less common in the real ratings situation. Extension of the RTM to the polytomous data, using the adjacent-category logits or cumulative logits, could expand the use of the RTM for the multiple ratings data and improve the generalizability of the findings from the current study. Another possible extension of the RTM can be to incorporate a longitudinal data structure in order to investigate effects such as rater drift over time. Although rater drift has been of interest for a couple of decades (e.g., Harik et al., 2009; Hoskens & Wilson, 2001), not many studies have explored rater drift using the multiple ratings data. Lastly, after fitting the model, it will be worthwhile to look at the rater-specific factor scores in detail. De Los Reyes (2011) argued that rating discrepancies across raters are substantively meaningful and should not be regarded simply as a type of measurement error. Rater-specific factor scores would be informative to monitor each rater’s scoring behavior as well as what needs to be strengthened in the rater training program.
Chapter 3.
Human Rater Monitoring with Automated Scoring Engines

3.1. Introduction

Interest in applying automated scoring to supplement or supplant human scoring has increased in recent years, and considerable effort has been directed toward researching and improving the automated scoring process (Attali & Burstein, 2006; Clauser, Kane, & Swanson, 2002; Landauer, Laham, & Foltz, 2003; Williamson, Xi, & Breyer, 2012). Little attention has been directed toward whether what we learn from automated scoring can be used to improve the human scoring process. In most current applications, automated scoring engines are calibrated using data from human scorers, so, clearly, obtaining ratings from humans that are the highest quality possible is paramount to the success of these automated scoring efforts. Additionally, the general public remains skeptical of the validity of automated scoring, so full implementation of automated scoring is increasing slowly. For the foreseeable future, human scoring will remain important; therefore, it is essential to continually improve the quality of the scores while making the scoring process as efficient as possible.

One common concern for those who manage and monitor human scoring projects is how to monitor rating quality in real time and over time (Myford & Wolfe, 2009). There are many potential procedures for doing this, but one that is used extensively is the process of having raters assign scores to validity papers. These papers are student responses that have been assigned consensus scores by expert raters before the operational scoring project. After the raters complete their training and begin operational scoring, validity papers are occasionally blindly seeded into the raters’ scoring queues, and the scores that the raters assign to the validity papers are recorded. When a rater has assigned scores to a sufficient number of validity papers, the scoring leaders review the scores, and compare them to the consensus scores assigned by experts. If large differences occur between the scores assigned by a particular rater and the scores assigned by experts, the scoring leaders may choose to provide feedback to that rater and/or take corrective actions.

This paper describes a study designed to evaluate the possibility of replacing human expert scores (HE) on validity papers with scores assigned by an automated scoring engine (AE). That is, rather than comparing the scores assigned by human raters to the consensus scores assigned by expert raters, we explore a scenario in which human rater scores are compared to scores derived from an automated scoring engine. The potential cost savings of implementing a rater monitoring system that relies on scores assigned by an automated scoring engine is considerable. A typical model for monitoring raters in an operational project includes convening a group of scoring leaders who select and assign scores to papers to construct validation sets. Although this process may be undertaken as part of the process of selecting training materials, this activity results in costs beyond those required to produce the
scores reported to students because convening that meeting may require the expert raters to travel and will also likely require paying the experts for their time. In addition, administering validity papers to raters during operational scoring projects adds additional cost beyond that required to assign a score to each student response. The validity papers typically need to be entered into scoring queues in a manner that prevents the assigned scores from being recorded as operational scores. In addition, the scoring distribution system must determine when and how frequently the validity responses are administered to each rater, and must then redirect the assigned scores to the rater monitoring system. Administering the validity sets to raters during operational scoring, generating reports that summarize the raters’ performance on the validity sets, and reviewing and then providing feedback to raters based on the information in the reports, introduce additional costs within a rating project beyond the costs associated with producing reported scores.

If validity paper administration could be replaced or reduced by the use of automated scores to monitor and evaluate raters, many of these costs could potentially be eliminated. Although automated scores would still require pulling papers to train the engine, assigning expert consensus scores would not necessarily be required for the papers. In addition, the entire process of having raters score responses that are not reported back to students for the sake of evaluating rater performance could be eliminated because the automated scores can be assigned to every operationally scored response. This would have the added benefit that every score assigned by a rater could be fed into the rater monitoring system. Currently, a very small number of validity papers is administered to each rater relative to the number of operational papers, due to the added cost of administering the validity papers. Thus, because the score of every response that a rater scores could be used to evaluate and monitor raters, that process could be considerably more precise than it is currently.

The purpose of the research reported here is to determine the effectiveness and efficiency of utilizing scores from automated scoring engines to monitor and provide feedback to human raters compared to the use of validity sets that are selected and assigned consensus scores by human scoring leaders. This study answers the question, “Are depictions of the quality of scores assigned by human raters comparable when monitoring is based on scores from an automated engine (AE) versus human experts (HE)?” If it were possible to utilize AE scores to monitor and provide feedback to human raters, the cost of the monitoring process in operational performance assessment scoring could be significantly reduced. To that end, we analyze the real data, taken from a rater straining study, applying mixed-effects ordered probit models (Rabe-Hesketh & Skrondal, 2012). Specifically, the analysis starts from a simple model that estimates a single rater effect (e.g., rater severity) and develops into a more complicated model that estimates multiple rater effects including the rater severity and inaccuracy. We then interpret how the model parameters and their transformations are related with the rater effects of our interest to illustrate the human rater monitoring using the AE for three aspects of the rater effects compared to the HE.
3.2. Methods

A body of literature describes how rater effects may be detected in rating data. In the family of Rasch modeling, the many-facet Rasch model (MFRM), which considers person, item, and rater facets, is a popular approach in the item response theory (IRT) modeling. In fact, the MFRM has the same mathematical form as the linear logistic test model (LLTM) (Fischer, 1973). The MFRM and the LLTM incorporate a rater severity parameter in an additive extension of the Rasch model. For example, based on the partial credit model (PCM; Masters, 1982) for the polytomous scores, the MFRM can be written as

\[
\text{logit}(P(X_{nir} = k | \theta_n, X_{nir} \in \{k, k-1\})) = \theta_n - \beta_i - \tau_{ik} - \rho_r
\]

where \(X_{nir}\) is the polytomous score among the \(k\) categories, given to examinee \(n\) on item \(i\) by rater \(r\), \(\theta_n\) is the latent proficiency of examinee \(n\), \(\beta_i\) is the difficulty of item \(i\), \(\tau_{ik}\) is the \(k\)th step difficulty for item \(i\), and \(\rho_r\) is the severity of rater \(r\).

In rater monitoring contexts, it is common to focus on the scores assigned to a single writing task even though it may be more efficient from a measurement perspective to base rater evaluations on scores assigned to several writing tasks. The reason for this is logistical—raters commonly are trained and assign scores to student responses using a single scoring rubric that was written for a specific prompt. Similarly, it is also common to have students respond to only a single prompt due to the amount of testing time required for open-ended assessment items. Therefore, we focus our attention to scoring contexts in which each rater assigns scores to student responses to a single prompt or item. In these contexts, the item facet in Equation (1.1) can be eliminated, which is analogous to the PCM as

\[
\text{logit}(P(X_{nir} = k | \theta_n, X_{nir} \in \{k, k-1\})) = \theta_n - \rho_r - \tau_{rk}
\]

where \(\tau_{rk}\) is the \(k\)th step severity for rater \(r\), and \(\rho_r\) is the overall severity of rater \(r\). Similar to the PCM, we can allow different step severity for each rater.

Although it appears the logit link has been more commonly used in the psychometric literature, we decide to use the probit link due to the software we choose for analysis, which are mixed effects ordered probit models. Furthermore, it should be noted that the human raters (HRs) under the training are qualitatively different from both HE and AE. In other words, the two types of target scores, HE and AE, are not exchangeable with the HRs (Raudenbush, 1993). For example, we have no a priori basis to predict how the parameter of human rater \(r\) will differ from that of another human rater \(r'\), but we may indeed have prior information about the parameter of HE and AE. Thus, in this study, considering the exchangeability principle, we apply two consecutive analyses: 1) HR + HE where we fix the model parameter for the HE and estimate them for individual HRs, and 2) HR + AE where we fix the model parameter for the AE and estimate them for individual HRs. If HE and AE depict the rater effects in the same way, we can expect that the parameter estimates for individual HRs from two consecutive analyses would show consistent and similar patterns.
In this section, we first introduce two models analyzed in this study, a mixed effects ordered probit model for rater severity, and another model for rater severity and rater-specific measurement error variances. We then relate the model parameters and their transformations to the rater effect indicators of our interest, which are rater severity, rater accuracy, and rater centrality. In summary, the total four types of analyses are conducted depending on the target score and model specifications, as summarized in Table 3-1.

Table 3-1. Four types of analyses

<table>
<thead>
<tr>
<th>Target Score</th>
<th>Mixed effects ordered probit models</th>
<th>Rater severity</th>
<th>Rater severity + Rater accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE (HR+HE)</td>
<td>Analysis I</td>
<td></td>
<td>Analysis II</td>
</tr>
<tr>
<td>AE (HR+AE)</td>
<td>Analysis III</td>
<td></td>
<td>Analysis IV</td>
</tr>
</tbody>
</table>

### 3.2.1. Mixed-effects ordered probit model for rater severity

We can specify models for ordinal scores by using either a generalized linear mixed model formulation or a latent-response formulation (Agresti, 2002; Rabe-Hesketh & Skrondal, 2012). There are three ingredients for a generalized linear mixed model formulation; link function (i.e., logit, probit), linear predictor (i.e., set of independent variables), and conditional distribution of the responses (i.e., multinomial distribution for ordinal responses).

First, we consider a cumulative ordinal probit model with a random intercept for person proficiencies, \( \theta_n \sim N(0, \psi) \). Mixed-effects ordered probit regression is ordered probit regression containing both fixed effects and random effects. In the absence of random effects, mixed-effects ordered probit regression reduces to ordered probit regression. The first model for the ordinal score \( X_{pr} \) assigned by rater \( r \) to person \( p \)’s essay is

\[
\Pr(X_{pr} > s \mid \theta_p) = \Phi(\theta_p - \kappa_s)
\]

where \( \Phi(\cdot) \) is the standard normal cumulative density function and \( \kappa_s \) is the threshold for score category \( s \). This model can also be written using the latent-response formulation, with the latent-response model and the threshold model specified as

\[
X_{pr}^* = \theta_p + \epsilon_{pr}, \quad \theta_p \sim N(0, \psi), \quad \epsilon_{pr} \mid \theta_p \sim N(0, \sigma)
\]

\[
X_{pr} = s \quad \text{if} \quad \kappa_{s-1} < X_{pr}^* < \kappa_s, \quad s = 1, \ldots, S
\]

respectively, with \( \kappa_0 = -\infty \) and \( \kappa_S = \infty \). This corresponds to a classical test theory model for \( X_{pr}^* \) when \( \theta_p \) represents the truth and \( \epsilon_{pr} \) represents measurement error. The model assumes that all raters \( r \) measure the same truth \( \theta_p \) with the same measurement error variance \( \sigma \) and assign scores using the same thresholds \( \kappa_s \) (\( s=1,\ldots,S-1 \)). We can allow for rater severities to be different by including rater-specific fixed-effects \( \rho_r \). However, one of the intercepts must
be set to zero to identify all the thresholds. Retaining the threshold model (1.5), we extend the latent-response model (1.4) to

$$X_{pr}^* = \theta_p + \rho_1 x_{1r} + \rho_2 x_{2r} + \cdots + \rho_{(R-1)} x_{(R-1)r} + \epsilon_{pr} \quad (1.6)$$

where $X_r = (x_{1r}, x_{2r}, \ldots, x_{(R-1)r})'$ are dummy variables for raters from 1 to $R-1$. This equation (1.6) corresponds to the Analysis I and Analysis III in Table 3-1 with use of different target scores $R$. The corresponding regression coefficients $(\rho_1, \rho_2, \ldots, \rho_{R-1})$ represent how much more generous or lenient each rater is than the last rater $R$. We arrange the last rater $R$ as the HE or AE, and fix the intercept of HE or AE as zero respectively in each analysis.

### 3.2.2. Mixed-effects ordered probit model for rater severity and rater-specific measurement error variances

Although the above model accommodates rater severity, it is relatively restrictive because it still assumes that all raters $r$ have the same measurement error variance $\sigma$. We can relax this homoskedasticity assumption by retaining the previous models (1.5) and (1.6) except that we now also allow each rater to have a rater-specific residual variance or measurement error variance $\sigma_r$, $\epsilon_{pr} | \theta_p \sim N(0, \sigma_r)$. This can be accomplished by specifying a linear model for the log standard deviation of the measurement errors using `gllamm` in Stata, the software we chose to use (Rabe-Hesketh & Skrondal, 2012):

$$\ln(\sqrt{\sigma_r}) = \ln(\sigma_r)/2 = \delta_1 x_{1r} + \delta_2 x_{2r} + \cdots + \delta_{(R-1)} x_{(R-1)r} \quad (1.7)$$

In this model for level-1 heteroscedasticity, we have again omitted the dummy variable for the last rater $R$ corresponding to HE or AE, which amounts to setting the standard deviation of the measurement error for this rater to 1 because $\exp(0)=1$. A constraint like this is necessary to identify the model because all thresholds $\kappa_s$ ($s=1, 2, 3$ in our data) are freely estimated. In terms of the above parameterization, the measurement error variance $\sigma_r$ for rater $r$ becomes $\exp(2\delta_r)$. In this model, each rater has his/her own mean and variance,

$$X_{pr}^* | \theta_p \sim N(\theta_p + \rho_r, \sigma_r), \quad \rho_R = 0,$$

but applies the same thresholds to the latent responses to generate the observed ratings. The cumulative probabilities are

$$\Pr(X_{pr} > s | \theta_p) = \Pr(X_{pr}^* > \kappa_s | \theta_p) = \Pr\left(\frac{X_{pr}^* - \theta_p - \rho_r}{\sqrt{\sigma_r}} > \frac{\kappa_s - \theta_p - \rho_r}{\sqrt{\sigma_r}}\right) = \Phi\left(\frac{\theta_p + \rho_r - \kappa_s}{\sqrt{\sigma_r}}\right).$$

This model can be thought of as a generalized linear model with a scaled probit link, where the scale parameter $\sqrt{\sigma_r}$ differs between raters, $r$. The covariate effect $\rho_r$ is constant across
categories, a property sometimes referred to as the parallel-regression assumption because the linear predictors for different categories are parallel.

### 3.2.3. Types of Rater Effects

*Rater effects* can be defined simply as “patterns of ratings that contain measurement error” (Wolfe & McVay, 2012). Raters may introduce errors into examinee scores in various ways—unfamiliarity with or inadequate training in the use of the rating scale, fatigue or lapses in attention, deficiencies in some areas of content knowledge that are relevant to making scoring decisions, or personal beliefs that conflict with the values adopted in the scoring rubric (Myford & Wolfe, 2003, 2004; Saal et al., 1980). In this study, we focus on three types of rater effects: severity/leniency, accuracy/inaccuracy, and centrality/dispersion.

Elsewhere, Wolfe and colleagues (Myford & Wolfe, 2004; Wolfe & McVay, 2012; Wolfe, 2004) provided a summary of rater effect indicators, including the three that we consider. However, those indicators are grounded and derived from a different modeling strategy, which was the Rasch rating scale model (Andrich, 1978) or the many-facet Rasch model (Linacre, 1994), and used the residuals after those models are fitted. Since the models analyzed in this study use different link function and different estimation methods, it is not yet shown whether those indicators are directly applicable. Instead, we identified rater effect indicators relating to the estimated model parameters and their formations. In this process, we utilize HE or AE as “target scores” by fixing their estimates and use them as the basis for making decisions about individual raters. We then compare the decisions that are made utilizing HE and AE scores as targets to determine whether HE and AE produce different depictions of the performance of individual human raters.

**Severity/Leniency**

We define “severity” as when a rater consistently assigns a lower score than the target scores. In contrast, “leniency” is defined as when a rater consistently assigns a higher score than the target scores. Commonly, severity/leniency has been evidenced by a decrease/increase in the average score associated with a rater (Wolfe, 2014). If severity/leniency exists in the scores, then some examinees will be incorrectly classified in decision making contexts such as during college admissions or placement or determining graduation qualification. Using the resulting estimates from the two sets of analyses that utilize HE and AE respectively, a rater was considered severe if his/her rater severity (\(\rho_r\)) estimate is significantly smaller than zero. Likewise, a rater was considered lenient if his/her severity estimate is significantly higher than zero.

**Accuracy/Inaccuracy**

A rater accuracy frame of reference portrays rating quality in terms of the deviation of a set of scores from another set of scores that is assumed to be valid (Wolfe, 2014). Accurate rating means that the rating is free from measurement error, and we consider the target scores, whether they are AE or HE, to be accurate ratings. Using the resulting estimates from the two
sets of analyses that utilize HE and AE respectively, a rater was considered inaccurate if his/her rater-specific measurement error variance ($\sigma_r$) estimate is significantly higher than the fixed value of HE or AE.

**Centrality/Extremity**

We define “centrality” as when a rater consistently assigns a score in the middle categories. The distribution of assigned scores can be compressed (centrality) or pushed into tails (extremity) when compared to true scores. For example, with four categories, a rater with centrality would likely to use 2 or 3 more often and use 1 or 4 less often than the target scores. When centrality/extremity exists in the scores, examinees in the tails of the distribution may be misclassified and/or decision makers may believe that examinees are less or more homogeneous than is actually the case. Commonly, centrality/extremity has been evidenced by a decrease/increase in the standard deviation of the scores associated with a rater (Wolfe, 2014). Using the resulting estimates from the two analyses, we calculate estimates of the “reduced-form” thresholds (Rabe-Hesketh & Skrondal, 2012).

$$\frac{\kappa_r}{\sqrt{\sigma_r}} = \begin{cases} 
\frac{\alpha_{r1} - \rho_r}{\exp(\delta_r)} & \text{for } r = 1 \\
\frac{\alpha_{r1} + \alpha_{rs} - \rho_r}{\exp(\delta_r)} & \text{for } r > 1 
\end{cases} \quad (1.8)$$

After transforming to the reduced-form thresholds, we decide which raters exhibit centrality using the thresholds of the HE and the AE as the basis. In detail, we compute the differences between the thresholds and compare those values to the corresponding differences computed from the HE and the AE. For example, if the gap between the thresholds for a certain rater is smaller than the corresponding gap from the HE, that rater is considered exhibiting centrality compared to the HE.

**3.2.4. Estimation**

Because our modeling requires us to handle the rater-specific heteroskedastic variances, `gllamm` command (Rabe-Hesketh, Skrondal, & Pickles, 2004) running in the widely available statistical package Stata (StataCorp., 2013) was used. Maximum likelihood estimation was implemented in the software `gllamm` using adaptive Gauss-Hermite quadrature with eight integral points for the mixed-effects ordered probit models (Rabe-Hesketh, Skrondal, & Pickles, 2005). Adaptive quadrature appears to be suitable when the posterior distribution is close to normal and when it is highly non-normal, whereas ordinary quadrature fails in the first situation (Rabe-Hesketh, Skrondal, & Pickles, 2002). Adaptive quadrature is computationally more efficient than ordinary quadrature and other computer intensive methods such as Markov chain Monte Carlo. It also provides a value for the maximized log likelihood useful for likelihood-ratio tests. In contrast to ordinary quadrature, adaptive quadrature also appears to give good parameter estimates for linear models, and is
useful for complex multilevel latent variable models that cannot yet be handled by other software, although computationally less efficient than other methods.

3.3. Illustration

3.3.1. Data

Each of 131 human raters (HRs) assigned holistic scores on a 4-point rating scale to 189 essays written by middle-school students in response to an explanatory prompt on a statewide writing assessment. Each essay was assigned a consensus “true score” by a panel of HE. More importantly, AE scores were obtained from an automated scoring engine that was calibrated on a separate sample of essays from the same population in this study. Figure 3-1 presents the structure of the empirical data used in this study. The data has a multilevel structure, in which the ratings are nested not only within students but also within each rater. Note again that we viewed the ratings from the HE and the AE as two structurally different target scores from the HRs and fixed their estimates, treating them as the criterion for other HRs.

Figure 3-1. Data Structure
3.3.2. Descriptive Results

Table 3-2 presents a direct comparison between the AE and the HE for the 189 essays. Out of the total 189 essays, the exact agreement rate between the HE and the AE was 64.02%. Fortunately, there was no case in which the discrepancy between the AE and the HE was 2 or more. When the off-diagonal elements were summed, the upper diagonal was 14.29% while the lower diagonal was 21.69%, which suggests that the AE was likely to produce lower scores compared to the HE in general. Interestingly, at the higher score category, the discrepancy was large: Only 0.53% when the AE gave 4 but HE gave 3, but 6.35% when the HE gave 4 but the AE gave 3.

Table 3-2. Direct comparison between the AE and the HE

<table>
<thead>
<tr>
<th></th>
<th>AE 1</th>
<th>AE 2</th>
<th>AE 3</th>
<th>AE 4</th>
<th>Total</th>
<th>Agreement %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>20</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>56</td>
<td>15</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>23</td>
<td>41</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>HE</td>
<td>1</td>
<td>20</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>56</td>
<td>15</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>23</td>
<td>41</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>10.58</td>
<td>5.29</td>
<td>0.53</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.17</td>
<td>29.63</td>
<td>7.94</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>12.17</td>
<td>21.69</td>
<td>0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>6.35</td>
<td>2.12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To get a sense of the rater effect indicators compared with the target scores, three types of corresponding descriptive statistics were calculated. First, regarding severity/leniency, the percentage of scores that was higher than the target scores given by human raters was calculated. From this aggregated information, the AE appears slightly more severe, and more human raters are classified as lenient when the AE was used as the target score, if we ignore the score categories.

Second, regarding accuracy/inaccuracy, the exact agreement rate with the target scores for each rater was calculated. When the HE was used as the target score, the exact agreement rates ranged from 38.10% to 73.55%, while the corresponding range was from 35.45% to 67.20% when the AE was used as the target score. The median rate was 56.61% for the HE and 53.97% for the AE. The correlation between the two sets of exact agreement rates was calculated as 0.72. Another way we examined the accuracy/inaccuracy was to calculate Spearman’s rank-based correlation between the rating from each rater and target scores. The range of the correlation was 0.42 to 0.77 when HE was used as a target score, and corresponding range was 0.39 to 0.77 when AE was used. Taken together, human raters appear to be classified similarly in terms of accuracy/inaccuracy when the HE score was used as the target score compared to when the AE score was used as the target score.

Third, regarding centrality/dispersion, the percentage the rater assigned score 2 or 3 was calculated for each rater. These values were 75.13% for the HE and 83.60% for the AE, which suggests that the AE showed more centrality than the HE in general. The range of the
corresponding values of 131 human raters was 42.33% to 88.89%. Among the 131 human raters, 53 raters had a higher percentage than the HE percentage, while only seven raters had a higher percentage than the AE percentage. These seven raters were the subset of the 53 raters. This aggregated information suggests that the AE would likely classify fewer raters as raters with centrality.

3.3.3. Results for the Mixed-effects Ordered Probit Models

Total four types of analyses, which include two different mixed-effects ordered probit models, one for rater severity and another for rater severity as well as rater-specific measurement error variances for two different data sets with different target scores, HR+HE and HR+AE, were conducted. First, based on the first model which specified only rater severity with the HE as a target score (Analysis I), thresholds were estimated as -1.65, 0.37, and 2.26 logit, and the variance of the person proficiency distribution ($\psi$) was estimated as 2.08. Second, based on the first model which specified only rater severity with the AE as a target score (Analysis III), thresholds were estimated as -1.51, 0.51, and 2.40 logit, and the variance of the person proficiency distribution ($\psi$) was estimated as 2.07. Although the threshold estimates are not directly comparable between Analysis I and Analysis III, the 95% confidence interval of the threshold estimates were overlapped each other. Overall, the thresholds estimates were consistently slightly higher when the AE was used as a target score. This implies that assuming the same measurement error variances across the raters, thresholds and variance of the person proficiency distribution were estimated quite similarly regardless we use either HE or AE as the target scores.

In contrast, when we allowed different measurement error variances across each rater in the second model (Analysis II and Analysis IV), the results showed different patterns to some extent. When the HE was used as a target score, the thresholds were estimated as -2.93, 0.63, 3.97, while the corresponding values were -2.19, 0.71, and 3.42 when the AE was used as a target score. Unlike the results from the Analysis I as well as Analysis III, the thresholds were shrunken toward when the AE was used as a target score. However, the 95% confidence intervals associated with each threshold were still overlapped each other. Moreover, the variance of the person distribution was estimated as 6.46 when the HE was used as a target score and the corresponding value was 4.26 when the AE was used as a target score. Compared with the results from the Analysis I as well as Analysis III, the variance estimates from both data sets became much larger, and the estimated variances were quite different. As the threshold estimates were spread wider when the HE was used as a target score, the person proficiency estimates were distributed in a wider range.

In order to test the model fit between the two mixed-effects ordered probit models analyzed in this study, we can use a likelihood-ratio test. Comparison between two models tests the null hypothesis that the measurement error variances are identical for the raters, against the alternative that the measurement error variances are different for at least two raters. Under the more restricted model (Analysis I and Analysis III), the measurement error variances are set to 1, and the thresholds of all raters are set to be equal; that is, in model (1.7),
the constraints $\delta_r = 0$ for $r=1,2,\ldots, R-1$ in place, but the intercepts $\rho_r$ for raters are free parameters. The more complex model is the same except that the constraints for $\delta_r$ are relaxed (Analysis II and Analysis IV). For both types of target scores, the likelihood-ratio test yielded that the more complex model fitted significantly better ($\chi^2(131)=1089.62, p<0.001$ for HR+HE and $\chi^2(131)=1074.11, p<0.001$ for HR+AE). AIC (Akaike Information Criterion; Akaike, 1974) that panelizes the complexity of the models also preferred the more complex model (41190.74 vs. 40363.11 for HR+HE, and 41213.14 vs. 40401.02 for HR+AE). This suggests strong evidence that at least two participating raters do not have same measurement error variances.

Next, to depict the rater effects in detail, we used the resulting estimates from the second model for each type of our rater effects indicators. Note again that in two different data sets, we fixed the parameters associated with target scores, HE and AE respectively. Beyond the statistical significance of each estimate, estimation of effect sizes for each type of our rater effects indicators needs to be considered. However, because the conventional effect size estimation is not appropriate and the associated standard errors are incorrect in multilevel structure setting (e.g., cluster randomized-trials, meta-analysis), we do not report the standard effect sizes at this moment (Donner & Klar, 2002; Rooney & Murray, 1996). Further work is needed on this topic.

**Severity/Leniency**

Figure 3-2 plots the rater-specific fixed effects ($\rho_r$), which correspond to the rater severity/leniency when the HE or the AE was used as a target score. In general, the rater-specific fixed effects when the AE was used as a target score tended to have higher values compared to when the HE was used as a target score, particularly at the lower range. Interestingly, the range of the estimates when the HE was used as a target score was wider from -2.43 to 2.19, while the range of the estimates when the AE was used as a target score was narrower from -1.78 to 1.98. This implies that when the AE was used as a target score, the differences in severity/leniency across the raters are likely to be condensed compared to when the HE was used as a target score.

To be specific, raters were classified as severe if the estimate is statistically significantly lower than zero, while lenient if the estimate is statistically significantly higher than zero. Based on this criterion, 48 raters were classified as lenient raters and 35 raters were classified as severe raters. Interestingly, classification of the raters in terms of the severity/leniency were exactly matched when the HE was used as a target score or when the AE was used as a target score. Taken together, although the magnitude of the estimate appears slightly different, more raters were classified as lenient while fewer raters were classified as severe regardless of the type of the target score. Thus, it appears reasonable to say that the AE depicts rater severity/leniency in the same way as the HE.
Figure 3-2. Severity/Leniency of each HR depending on the target score

*Accuracy/Inaccuracy*

Figure 3-3 illustrates the estimates of the rater-specific measurement error variances ($\sigma_r$) depending on the target score. Again, estimates for the HE and the AE were fixed as 1 respectively (Rabe-Hesketh & Skrondal, 2012). In general, most of the human raters under the training demonstrated higher measurement error variances compared to both types of target scores. In terms of the magnitude, the measurement error variances were estimated higher across all the raters when the HE was used as a target score, compared to when the AE was used as a target score. However, given that the estimates of rater-specific measurement error variances are located in a quite higher range, using the fixed value 1 as the criterion appears very stringent. In detail, the measurement error variances were distributed wider when the HE was used as a target score ranging from 0.98 to 8.48, while the corresponding values when the AE was used as a target score ranged from 0.66 to 5.64.

Although comparison to the fixed value 1 seems too harsh for most of the raters, raters were classified as inaccurate if the estimate is significantly higher than the fixed value 1. Using the standard error associated with the rater-specific measurement error variance estimates, 95% confidence intervals for each rater were constructed and compared with the fixed value 1. Based on this criterion, as expected, most of the raters, 122 out of 131 raters, were classified as inaccurate when the HE was used as the target score. In contrast, 87 out of 131 raters were classified as inaccurate when the AE was used as the target score. These 87
raters were the subset of the 122 raters labeled as inaccurate based on the HE. Taken together, considering the limitation that we used the harsh criterion, it is not easy to say that the AE provides the similar picture as the HE for rater accuracy, by labeling fewer raters as inaccurate.

![Figure 3-3. Rater-specific measurement error variance each HR depending on the target score](image)

**Centrality/Extremity**

**Error! Reference source not found.** displays reduced-form thresholds, red for the first threshold, orange for the second threshold, and green for the third threshold, transformed using the Equation (1.8). Solid line represents the thresholds when the HE was used as a target score, and the dotted line represents the thresholds when the AE was used as a target score. The thresholds for each rater are indexed as larger hollow circle when the HE was used as a target score, and smaller solid circles are used when the AE was used as a target score. As illustrated, the locations of two types of the circles were quite similar apparently due to the parallel-regression assumption. However, as discussed before, the thresholds were shrunken when the AE was used as a target score compared to when the HE was used as a target score, and this would likely affect on our decision on rater centrality/extremity.

Specifically, we computed the differences between the thresholds and compared those values to the corresponding differences from each threshold when HE and AE were used as target scores. As the shrunken range of the AE indicated, the differences between the thresholds were larger when the HE was used as a target score (3.56 between the first and the
second threshold, and 3.33 between the second and the third threshold), compared to when the AE was used as a target score (2.90 between the first and the second threshold, and 2.72 between the second and the third threshold). In detail, when the HE was used as a target score, the gap between the first and the second threshold of human raters ranged from 1.22 to 3.59 with the median of 2.13. The gap between the second and the third threshold ranged from 1.15 to 3.36 with the median of 2.00. The corresponding ranges when the AE was used as a target score showed quite similar ranges: the gap between the first and the second threshold ranged from 1.22 to 3.58 with the median of 2.13, and the gap between the second and the third threshold ranged from 1.14 to 3.35 with the median of 2.00. Taken together, the ranges from the human raters were much smaller than the ranges computed from the HE and the AE, implying that this criterion looks very stringent.

Given that the thresholds of each rater based on either HE or AE yielded similar values, more raters are likely to be classified as demonstrating centrality when the HE was used as a target score. Based on this criterion, 130 out of 131 raters were classified as exhibiting centrality when the HE was used as a target score, while 122 out of 131 raters were classified as exhibiting centrality when the AE was used as a target score. These 122 raters were the subset of the 130 raters labeled as central based on the HE. Overall, considering the limitation that the used criterion may be too strict, it appears that the AE depicts rater centrality slightly different from the HE, by labeling fewer raters as centrality.

3.4. Conclusion and Discussion

Since interest in implementing automated scoring systems is increasing, it is important to examine whether AE scores can be used to monitor and provide feedback to human raters instead of expensive HE scores. In order to explore this consequential validity of the AE scores in rater monitoring, we analyzed an empirical rater training data using two different mixed-effects ordered probit models. Applying the mixed-effects ordered probit models to the data where ratings from many human raters along with the target scores (i.e., AE or HE) would guide us to compare how raters behave compared to the target scores. A simpler model incorporated only the rater severity and a more complicated second model also allowed level-1 heteroscedasticity for rater-specific measurement error variances. Comparison between these two models revealed that it is necessary to relax the same measurement error variances across the raters.

In particular, the modeling strategy used in this study is convenient and straightforward since it estimates the model parameters. Those model parameters can be directly related with types of rater effects indicators of our interest, such as rater severity/leniency and rater accuracy/inaccuracy. Fixed-effects rater location estimates were used for rater severity/leniency, and rater-specific measurement error variances were used for rater accuracy/inaccuracy. We also computed the thresholds using the resulting estimates to depict rater centrality/extremity. For the data in this study, the results showed that the AE depicts the HE similarly in terms of the rater severity, while slightly different in terms of the rater accuracy and rater centrality. In particular, the AE labeled the same raters as lenient and
severe as the HE did. However, AE classified fewer raters as demonstrating inaccuracy and centrality. One thing to note here is that employed criteria for rater inaccuracy and centrality appeared too stringent. We fixed the model parameters associated with the HE and the AE and used them as the criteria, but the range of the parameter estimates were quite large compared to those fixed values. This may describe the situation well, but this might be also related with the relatively smaller sample size compared to the huge number of model parameters, which led to larger standard errors. Thus, for future study, it may be interesting to explore the use of these models to larger data sets to examine the findings in a similar context, or to experiment with more reasonable and realistic criteria for rater inaccuracy and rater centrality. Furthermore, it will be interesting to estimate effect sizes for each type of our rater effect indicators considering the multilevel ordinal data in this study. Recently, Hedges (2007) defined and proposed effect size estimation in cluster randomized-trials by correcting the intraclass correlation. Larsen and Merlo (2005) also suggested calculation of the median odds ratio as a measure of heterogeneity, which can also be understood as a type of effect size. However, they were not designed for the ordinal data, and has not yet been shown whether their methods can be directly applicable for the multilevel probit regression we used in this study. Thus, it will be worthwhile to examine and possibly modify their methods for our setting in order to develop more practical criteria and utilize them in addition to the statistical significance.

In addition, we illustrated the use of two different mixed-effects ordered probit models with an empirical example. Both models assumed the parallel-regression assumption that the linear predictors for different categories are parallel. Given the sample size and the number of model parameters, we were not successful in relaxing this assumption. However, relaxation of this assumption would lead to estimation of model parameters that are more directly related with the rater centrality/extremity, by allowing rater-specific thresholds.
Figure 3.3. Thresholds compared to the HE (solid) and the AE (dotted)
Chapter 4.
Structured Constructs Models (SCM) based on the Change-point Analysis

4.1. Introduction

Learning progressions describe sophisticated ways of thinking about a topic that can be utilized as children learn about and investigate a topic (National Research Council, 2007). Learning progressions enable us to illustrate and to test our theories about the conceptual pathways that students follow while learning in a domain. One example in which the core ideas of learning progressions are embodied regarding curriculum development and assessment is a construct map approach (Wilson, 2005). A construct indicates an underlying latent variable that is assumed to be continuous and to range from one extreme to another, and a construct map is an ordering of locations along the continuum that define qualitatively different levels of performance or competence on the construct. The levels of the learning progression relate to the levels of the construct maps, and the descriptions of the levels exhibit an aspect of the development. As the levels move upwards, the sophistication increases from level to level.

One of the complications in real situations, however, is that there are often complex multiple constructs within a learning progression. In that case, a multidimensional item response model can be one of the appropriate measurement models (Adams, Wilson, & Wang, 1997; Reckase, 2009). Multiple latent ability dimensions are incorporated into multidimensional item response modeling, and the extent to which two or more dimensions are correlated among constructs within a learning progression can be explored. Moreover, one promising aspect of the structure of learning progressions is the possibility of specifying relationships that go beyond the overall correlation of two or more dimensions. The detailed specification of the inner structure of a learning progression allows for the formulation of hypotheses regarding potential relationships that may exist between the interim levels of development across a set of dimensions. In many circumstances, it is reasonable to hypothesize that reaching a certain level of one dimension (i.e., a required dimension) boosts a probability of reaching a certain level of a second dimension (i.e., a target dimension).

Figure 4-1 illustrates this relationship in the motivating example, which concentrates on the two constructs from the Assessing Data Modeling and Statistical Reasoning (ADM) Project, Conceptions of Statistics (CoS) and Chance (Cha) (Lehrer & Wilson, 2011). The CoS construct describes how students develop their concepts of the meaning and the uses of statistics by informally describing a distribution using shapes to understand statistics as a measure of summarizing a sampling distribution. The Cha construct represents students’ progression in understanding probability as a measure of uncertainty. According to the learning progression, students are expected to understand that chance yields a distribution of outcomes as they progress to more sophisticated levels in the Cha construct. In this project,
we see each successive pair of levels in CoS and Cha to be a requirement levels and target levels respectively. As shown, arrows connect specific levels of the two constructs of the theoretical ADM learning progression. A link from CoS level 3 to Cha level 3 represents that a student is more likely to reach level 3 on the Cha construct if he/she has reached level 3 on the CoS construct. Likewise, a link from Cha level 6 to CoS level 4 represents that a student is more likely to reach level 4 on the CoS construct if he/she has reached level 6 on the Cha construct.

The structured constructs model (SCM) was designed for just this sort of situation (Wilson, 2009, 2012). In a previous study, the SCM was formulated by building on ordered latent class models (Diakow, Irribara, & Wilson, 2011). Following the ordered latent class models formulation (Croon, 1990), latent classes are ordered from low to high, and the probability of a correct answer for the items increases as examinees progress from lower classes to higher classes. The study by Diakow and collegues (2011) extended the ordered latent class models into multiple latent variables in which links between constructs were expressed as the joint probabilities of class membership. For example, being on one level of a required dimension makes it viable that a person would be in a particular level of the target dimension. Here, the underlying model for each construct is seen as a series of latent classes rather than a latent continuum, and the persons within the same latent class have the same probabilities of answering a set of items correctly and, thus, the same ability or proficiency.

However, it is plausible that students within a given level of the dimension may vary in their overall proficiency within that level (e.g., between-category qualitative difference and within-category heterogeneity in De Boeck, Wilson, & Acton (2005)). This possibility was explored in Choi (2013) by placing judgmental cut scores along a latent continuum to identify the levels of the construct. The hypothesized links, such as respondents in one level of the
required dimension being more likely to belong to a particular level of the target dimension, are incorporated by the use of discontinuity parameters.

In line with Choi (2013)’s study, this study presents a modification of the approach proposed in Choi (2013) in two ways by introducing the idea of change-point analysis. First, instead of placing pre-determined or ad-hoc cut scores, the cut scores, which correspond to the change points in this study, will be directly estimated. Second, beyond the Two Level Cases (e.g., mastery vs. non-mastery), this chapter extends the application of the SCM to multiple levels, such as Three Level Cases (e.g., Level 1, Level 2, and Level 3), which occur more frequently in the learning progressions framework. Thus, the main goal of this study is to describe and propose one possible way to formulate a measurement model for complicated learning progressions using the SCM approach based on change-point analysis, which can be useful to estimate the cut scores and discontinuities at the cut scores. This chapter briefly describes previous statistical frameworks for the SCM and presents an SCM based on the change-point analysis. Subsequently, the results of applying the method to the simulated data and the empirical data are presented and discussed. Lastly, the chapter concludes with suggestions for further research.

4.2. Structured Constructs Models

The SCM idea provides a theoretical framework of measurement models for complicated learning progressions in which multiple constructs are involved and the relations among the levels of the constructs are hypothesized in advance. Therefore, the three most important features in SCM modeling are (1) to identify the cut score(s) on the latent requirement construct continuum, (2) to classify the examinees into the levels on the requirement construct, and (3) to model the links so that the hypothesis on the specific relationships between the requirement levels and target levels can be incorporated. In this section, the SCM approach based on the change-point analysis that satisfies those features is proposed after reviewing the previous statistical frameworks for the SCM.

4.2.1. Previous Statistical Framework of Structured Constructs Models

The previous statistical frameworks for the SCM can be divided into two approaches: the SCM based on the ordered latent class models and the SCM for the latent continuous variables. First, the latent class analysis primarily aims to identify subgroups of examinees by relating the observed item responses to a set of latent discrete variables. These categorical latent variables indicate the class membership of the examinees, which are mutually exclusive and exhaustive to one another. After identifying the latent groups, each examinee is classified into one of the levels (classes) according to his or her latent class membership (Maris, 1999). Considering the ascending structure of the learning progressions from lower levels to higher levels, the ordered latent class model is considered to be a way to describe learning progressions. Diakow et al. (2011) extended the ordered latent class models to two latent variables and applied it as the SCM approach. Specifically, the crossed levels of the two constructs were considered to be classes, and the links between the constructs were expressed
using the joint probabilities of class membership. In their approach, the levels were identified with latent discrete variable and the underlying person proficiency was not assumed as the latent continuum. Instead, examinees within the same level were assumed to be identical regarding their probabilities of providing particular responses to the items.

However, it is plausible that students within a given level of the dimension may vary in their overall proficiency within that level (e.g., between-category qualitative difference and within-category heterogeneity in De Boeck, Wilson, & Acton (2005)). As the second approach, Choi (2013) proposed an SCM for the latent continuous variables for a simple case in which there is a single connection between two constructs and the item responses are dichotomous. Each construct is assumed to be continuous and to be composed of two levels (e.g., master versus non-master or proficiency versus non-proficiency) as illustrated in Figure 4-2 (a). The heterogeneity in construct level is represented with a normal distribution for each construct, although normality of the distributions is not required. The hypothesized link represents that if person \( p \) is considered to be at the proficient level on the requirement construct, he/she is more likely to be classified in the proficient level on the target construct. This hypothesized link between the levels is expressed as an arrow from the proficient level in the requirement construct to the proficient level in the target construct. The construct from which the arrow initiates is referred to as the “requirement construct,” and the construct at which the arrow terminates is referred to as the “target construct,” and \( \theta_{pR} \) and \( \theta_{pT} \) denote the continuous latent variable of person \( p \)’s proficiency in the requirement and target constructs, respectively.

The link between the levels of the two constructs can be modeled by the introduction of a discontinuity parameter, which is similar formally to the saltus parameter (Draney & Wilson, 2007; Wilson, 1989). In the saltus model, as an individual progresses from the lower levels to the higher levels, a sudden spurt or change occurs. Specifically, in the saltus model, the classes of the persons, to be estimated, represent different development stages or levels,
and the groups of items are specified to allow persons at or above the developmental stage to have the advantage of being able to give better answers in that stage. The saltus parameter, \( \tau_{ck} \), quantifies these discontinuities as additive effects on the item parameters of all items in item group \( k \) when people in group \( c \) respond to those items. Similarly, in the SCM approach proposed by Choi (2013), the hypothesized links between the two constructs are assumed to induce the discontinuities. Particularly in the requirement construct, the probability that person \( p \) gives a correct response on item \( i \) is written according to the Rasch model (Rasch, 1960) as:

\[
\Pr(Y_{piR} = 1 | \theta_{pR}) = \frac{\exp(\theta_{pR} - \beta_{iR})}{1 + \exp(\theta_{pR} - \beta_{iR})}. \tag{1.1}
\]

On the contrary, the probability of success on the items in the target construct depends on the ability in the target construct \( \theta_{pT} \) as well as the ability in the requirement construct \( \theta_{pR} \). More specifically, if \( \theta_{pR} \geq C_1 \), where \( C_1 \) is the cut score in the requirement construct, the probability is augmented by \( \lambda_1 \), and if \( \theta_{pR} < C_1 \), then the probability is augmented by \( \lambda_2 \). In other words, two discontinuity parameters, \( \lambda_1 \) and \( \lambda_2 \), can be considered to be advantage (or disadvantage) parameters in the target construct according to the level in the requirement construct. Then, the probability of person \( p \)’s correct answer on item \( i \) in the target construct is expressed as:

\[
\Pr(Y_{piT} = 1 | \theta_{pT}, \theta_{pR}) = \frac{\exp(\theta_{pT} + \lambda_1 f(\theta_{pR}) + \lambda_2 (1 - f(\theta_{pR})) - \beta_{iT})}{1 + \exp(\theta_{pT} + \lambda_1 f(\theta_{pR}) + \lambda_2 (1 - f(\theta_{pR})) - \beta_{iT})}, \tag{1.2}
\]

where \( f(\theta_{pR}) = 1 \) when \( \theta_{pR} \geq C_1 \) and \( f(\theta_{pR}) = 0 \). In this formulation, \( \beta_{iT} (i = 1, \ldots, I_T) \) indicates the difficulty of the \( i \)th item in the requirement construct, and \( \beta_{iT} (i = 1, \ldots, I_T) \) is the difficulty of the \( i \)th item in the target construct. In Choi’s (2013) study, the cut score \( C_1 \) is predetermined as the mean of the item difficulties \( (C_1=0) \) instead of being estimated directly.

### 4.2.2. Structured Constructs Model based on the Change-Point Analysis

\[
y_i = \beta_0 + \beta_1 x_i + \beta_2 I(t \geq \pi) + \beta_3 I(t \geq \pi) x_i,
\]

The change-point analysis has been applied in various fields to detect the number of change (discontinuity) points and their locations (e.g., Chen & Gupta, 2011; Chib, 1998; Hall, Lipton, Sliwinski, & Stewart, 2000; Hinkley, 1970; Western & Kleykamp, 2004). Traditionally, the change-point analysis has been used for time-series data to detect the significant radical change(s) during some time period and for medical data to find the location of a certain gene that shows an abrupt difference. For an example of time-series data, a simple regression model that embodies the change-point hypothesis for a continuous dependent variable is written as:

\[
\hat{y}_t = \beta_0 + \beta_1 x_i + \beta_2 I_t(\pi) + \beta_3 I_t(\pi) x_i, \quad t = 1, \ldots, T, \tag{1.3}
\]
where for the change-point indicator \( I_t(\pi) = 0 \) for \( t < \pi \) and \( I_t(\pi) = 1 \) for \( t \geq \pi \), \( y_t \) are the observations of the dependent variable at time point \( t \), \( x \) is a covariate, and \( \beta \)'s are the regression coefficients. Apart from the specified covariates, the primary purpose of this model is to estimate the time-point in which the change occurs (\( \pi \)) as well as the mean change in the dependent variable at that change-point (\( \beta_2 \)).

In this study, the focus is quite similar to this simple form of change-point analysis: the estimation of the change point (\( \pi \)) and the average magnitude of abrupt change at the change-point (\( \beta_2 \)). However, rather than time \( t \), we are interested in a change point in ability (\( \theta \)). Specifically, we aim to estimate the change point that corresponds to the cut scores in the requirement construct, which leads to the classification of the examinees according to the cut scores. We also aim to estimate the magnitude of abrupt change at the change-point, which corresponds to the discontinuity at the cut scores that incorporate the mean change in the latent ability distribution in the target construct according to the hypothesized links. However, our case is more complicated in terms of the relationship between the unit of analysis and the scale of the model parameters. In many applications of the change-point analysis, the dependent variable (e.g., wage growth) is measured or adjusted in the unit of the change-point (i.e., time, for example in years). On the other hand, our unit of analysis (e.g., item-level response data) is different from the scale of the change-points, which will be in units of logit. We use the item-level response data, which are commonly dichotomously or polytomously scored, and we aim to estimate both the change point and the discontinuity on the continuous scale, along with the item parameters. Because of the properties of the Rasch model family we use in this study, the item parameters and the person latent abilities were scaled on the same logit scale (Rasch, 1960; von Davier & Carstensen, 2007). This allows us to estimate the cut scores along with the item parameters and to formulate the discontinuity parameters at the estimated cut scores on the same scale as the latent abilities of persons.

Based on the above similarities and complexities, we refer to the proposed model as a change-point structured constructs model (change-point SCM). In this section, the change-point SCM is considered in two forms: (a) when there are two levels for each construct (Two Level Case, e.g., proficient vs. non-proficient), and (b) when there are multiple levels for each construct (Multiple Level Case, e.g., Level 1, Level 2, and Level 3). In both cases, each item is designed to measure only one construct as in the between-item multidimensional item response model (Adams et al., 1997).

**Learning progressions with two levels (Two Level Case)**

In this study, the hypothesized link between the levels of the two constructs is modeled based on a change-point analysis to allow for the estimation of the cut score and the discontinuity at the cut score. Briefly, we propose the formulation that the cut score is determined by the overall item difficulty in the requirement construct. In the original Rasch model in Equation (1.1), item parameters (\( \beta_{ir} \)), namely, item difficulties, are estimated for individual items. Alternatively in this study, the item parameters are re-parameterized as the combination of the average difficulty across all items and the deviation from the average
difficulty for each item. Thus, the re-parameterized Rasch model is written as follows and used for the requirement construct:

\[
\Pr(Y_{pR} = 1 | \theta_{pR}) = \frac{\exp(\theta_{pR} - (\gamma + \tau_{iR}))}{1 + \exp(\theta_{pR} - (\gamma + \tau_{iR}))},
\]

(1.4)

where \( \gamma \) represents the intercept that corresponds to the average difficulty of all items. This estimate serves as the directly estimated cut score, and \( \tau_{iR} (i = 1, \ldots, I_R) \) indicates the deviation from this average difficulty for each item \( i \) in the requirement construct. Using the estimated cut score \( \gamma \) in the requirement construct, examinees are separated into the proficient level and the non-proficient level on the requirement construct. Given that the estimated cut score (\( \gamma \)) corresponds to the change-point (\( \pi \)) in Equation (1.3), the probability equation for the target construct is now written as:

\[
\Pr(Y_{pT} = 1 | \theta_{pT}, \theta_{pR}) = \frac{\exp(\theta_{pT} + \lambda I(\theta_{pR} \geq \gamma) - \beta_{iT})}{1 + \exp(\theta_{pT} + \lambda I(\theta_{pR} \geq \gamma) - \beta_{iT})},
\]

(1.5)

where \( I \) is the indicator function that returns 1 when \( \theta_{pR} \geq \gamma \) and is 0 otherwise, and \( \lambda \) represents the discontinuity parameter according to the level in the requirement construct. In particular, for the persons in the proficient level on the requirement construct, if \( \theta_{pR} \geq \gamma \), the probability is augmented by \( \lambda \), which represents the size of the jump at the estimated cut score \( \gamma \). In the target construct, \( \beta_{iT} (i = 1, \ldots, I_T) \) indicates individual item difficulties.

Equations (1.4) for the requirement construct and (1.5) for the target construct constitute the change point SCM for the dichotomous data and learning progressions with two levels (Two Level Case). Unlike the formulations in Choi (2013) that estimated the two discontinuity parameters at the pre-determined cut score, a single discontinuity parameter \( \lambda \) for those who had higher proficiencies than the estimated cut score in the requirement construct was estimated. This corresponds with the magnitude of mean change (\( \beta_2 \)) in Equation (1.3). In other words, \( \lambda \) represents the size of the jump to be estimated as the magnitude of the average increase on the logit scale for those who exceeded the cut score in the requirement construct. Moreover, as in the Equation (1.4), if the discontinuity parameter \( \lambda \) in the target construct is assumed to be zero while the cut score \( \gamma \) in the requirement construct is still estimated, we call it as the re-parameterized multidimensional Rasch model (RMRM). In the context of the multidimensional assessments, the RMRM can be useful if one is interested in estimating the cut score in one dimension without a particular hypothesis on the specific relationships between the levels. In this study, the RMRM is used as the baseline model for the Two Level Case to investigate the consequences of ignoring the discontinuity parameters.

As in the multidimensional Rasch models, \( \theta_{pR} \) and \( \theta_{pT} \) are assumed to follow a bivariate normal distribution:

\[
\begin{bmatrix}
\theta_{pR} \\
\theta_{pT}
\end{bmatrix}
\sim
\text{MVN}
\begin{bmatrix}
\mu_R \\
\mu_T
\end{bmatrix},
\begin{bmatrix}
\sigma^2_R & \sigma_{RT} \\
\sigma_{RT} & \sigma^2_T
\end{bmatrix}
\]

For identification purposes, the means of the bivariate normal distributions are set to \( \mu_R = \mu_T = 0 \), and the sum of the item
difficulties in the target construct $\beta_i (i = 1, \ldots, I_T)$ is constrained to be zero (but the difficulties of the requirement construct are not constrained). The cut score parameter, the discontinuity parameter, and the deviations from the cut score for each item $i$ in the requirement construct are freely estimated.

Figure 4-3 plots an example of the proficiency distributions in the requirement construct and the target construct based on the change-point SCM for the Two Level Case. There are two distributions in the target construct: a black line for the non-proficient group and a red line for the proficient group. The heterogeneity in each group is represented with a normal distribution for each construct, although normality of the distributions is not required. As shown in Equation (1.4), the examinees are divided into a proficient group and a non-proficient group based on the estimated cut score parameter, $\gamma$. Based on this estimated cut score in the requirement construct, proficient examinees in the requirement construct are more likely to achieve a proficient level in the target construct. This advantage is reflected in the discontinuity parameter $\lambda$ in Equation (1.5) that increases the probability of responding to items correctly in the target dimension for the proficient examinees. As shown in Figure 4-3, for the distribution of the proficient group, the red line moves up on average as much as the magnitude of the discontinuity parameter. According to the hypothesized link between the two constructs, the discontinuity parameter is most likely to have a positive value. Otherwise, the discontinuity parameter would decrease the probabilities in the target construct, which can
happen in the formal mathematical model, but it contradicts the eventual hypothesis of the learning progression.

**Learning progressions with multiple levels (Multiple Level Case)**

In real assessments, learning progressions are more likely to have more than two levels. For example, when there are three levels in the learning progressions (e.g., Level 1, Level 2, Level 3), there can possibly be two hypothesized links between the two constructs as illustrated in Figure 4-2 (b). If person \( p \) is considered to be on Level 2 in the requirement construct, he/she is more likely to reach Level 2 on the target construct, and the equivalent holds for the persons on Level 3. Likewise, these hypothesized links between the two constructs are expressed as arrows from Level 2 in the requirement construct to Level 2 in the target construct, and same is true for Level 3 in the requirement construct and Level 3 in the target construct.

When the learning progressions consist of three levels for each construct and the items are designed to have three score categories, a similar concept of re-parameterization can be applied to the partial credit model, which is a family of the Rasch model for the polytomous data (PCM; Masters, 1982). Originally, as written in Equation (1.6), the PCM specifies the probability of responding in the \( j \)th score category of item \( i \) for person \( p \) as a function of the person ability \( \theta_p \) and item-specific step parameters \( \delta_{ij} \) (\( j>0 \)).

\[
\text{Pr}(Y_{pi} = j | \theta_p) = \frac{\exp[\sum_{l=0}^{j} (\theta_p - \delta_{il})]}{\sum_{k=0}^{m_i} \exp[\sum_{l=0}^{k} (\theta_p - \delta_{il})]} \quad j = 0, 1, \ldots, m_i
\]  

(1.6)

Alternatively, the step parameters \( \delta_{ij} \) for individual items can be re-parameterized as the combination of the overall step difficulties \( \gamma_j, j>0 \) and the deviations from the overall step difficulties for each item \( (\tau_{ij}, j>0) \). For example, if there are three score categories \( (j=0,1,2) \), two overall step difficulties \( \gamma_1 \) and \( \gamma_2 \) will be estimated. These overall step difficulties can be considered as the estimated cut scores between Level 1 and Level 2 and Level 2 and Level 3, respectively. In other words, \( \gamma_1 \) represents the overall step difficulty to advance from Level 1 to Level 2, and \( \gamma_2 \) represents the overall step difficulty to progress from Level 2 to Level 3. Then, \( \tau_{il} (i = 1, \ldots, I) \) indicates the deviation from the first cut score for each item \( i \), and \( \tau_{l2} (i = 1, \ldots, I) \) indicates the deviation from the second cut score for each item \( i \). Then the probability equation for the requirement construct is now written as:

\[
\text{Pr}(Y_{pIR} = j | \theta_{pR}) = \frac{\exp[\sum_{l=0}^{j} (\theta_{pR} - (\gamma_l + \tau_{il}))]}{\sum_{k=0}^{m_i} \exp[\sum_{l=0}^{k} (\theta_{pR} - (\gamma_l + \tau_{il}))]}, \quad j = 0, 1, \ldots, m_i
\]  

(1.7)

Based on the estimated cut scores \( \gamma_1 \) and \( \gamma_2 \), examinees are separated into the levels of performance on the requirement construct. Then, the probability of getting score \( j \) on the items in the target construct depends on the ability in the target construct \( \theta_{pT} \) as well as the ability in the requirement construct, \( \theta_{pR} \). By introducing the discontinuity parameter for the
hypothesized link, the probability of person $p$ receiving the score category $j$ on item $i$ in the target construct is expressed as:

$$
\Pr(Y_{pT} = j | \theta_{pR}, \theta_{pT}) = \frac{\exp[\sum_{t=0}^j (\theta_{pT} - \delta_{it}^T + \lambda_j \mathbb{1}[\theta_{pR} \geq \gamma_j])] \sum_{k=0}^m \exp[\sum_{t=0}^k (\theta_{pT} - \delta_{it}^T + \lambda_j \mathbb{1}[\theta_{pR} \geq \gamma_j])]}{\sum_{k=0}^m \exp[\sum_{t=0}^k (\theta_{pT} - \delta_{it}^T + \lambda_j \mathbb{1}[\theta_{pR} \geq \gamma_j])]}, \quad j = 0, 1, \ldots, m_i, \quad (1.8)
$$

Similar to the example for the Two Level Case, $I$ is the indicator function that returns 1 when $\theta_{pR} \geq \gamma_j$ and it is 0 otherwise. For the persons in Level 2 on the requirement construct, if $\gamma_2 > \theta_{pR} \geq \gamma_1$, the probability is augmented by $\lambda_1$, which represents the size of the jump at the estimated cut score $\gamma_1$. In the same manner, for the persons on Level 3 of the requirement construct, if $\theta_{pR} \geq \gamma_2$, the probability in the target construct is increased as much as $\lambda_2$. At each estimated cut score, the advantage can manifest itself as different “boosts”. This is shown by the two discontinuity parameters in Equation (1.8). However, it is possible to impose various constraints on the two discontinuity parameters, such as an equivalent assumption ($\lambda_1 = \lambda_2$), based on the theoretical expectations on the magnitude of discontinuity. This possibility will be examined with two empirical examples as a way of empirical validation of the hypothesized links. In the target construct, analogous to the standard PCM, the item parameter $(\delta_{ijT})$ can be decomposed into the overall item difficulty for $i$ ($\beta_{ijT}$) and item-specific step difficulties ($\tau_{ijT}$).

Equations (1.7) for the requirement construct and (1.8) for the target construct constitute the change point SCM for the Multiple Level Case with the polytomous data. This formulation can be extended straightforward to the situations in which more than three levels of performance exist in the learning progression by increasing the number of cut scores and the discontinuity parameters. If it is assumed that there is no discontinuity in the target construct ($\lambda_j = 0$) while estimating the cut scores ($\gamma_j$) is still of interest, we call it as the re-parameterized multidimensional PCM (RMPCM). The RMPCM can be used when the multidimensional assessments are composed of the multiple levels and one is interested in estimating the cut scores to classify the examinees in one dimension without particular knowledge on the specific relationships with other constructs. In this study, the RMPCM is used as the baseline model for the Three Level Case to investigate the consequences of ignoring the discontinuity parameters.

As in the standard multidimensional PCM, $\theta_{pR}$ and $\theta_{pT}$ are assumed to follow a bivariate normal distribution:

$$
\begin{bmatrix}
\theta_{pR} \\
\theta_{pT}
\end{bmatrix} \sim \text{MVN}
\begin{bmatrix}
\mu_R \\
\mu_T
\end{bmatrix},
\begin{bmatrix}
\sigma^2_R & \sigma_{RT} \\
\sigma_{RT} & \sigma^2_T
\end{bmatrix}.
$$

For identification purposes, a higher cut score parameter is constrained to be the negative of another cut score parameter ($\gamma_2 = -\gamma_1$), and the means of bivariate normal distribution are set to $\mu_R = \mu_T = 0$. In the target construct, the sum of the overall item difficulties $\beta_{ijT} (i = 1, \ldots, I_T)$ and the sum of the step difficulties within each item $i$ ($\tau_{ijT}$) are constrained to be zero. For the requirement construct, the sum of the deviations $\tau_{ijR} (i = 1, \ldots, I_R)$ within each item is also constrained to be zero.

58
For the Three Level Case, Figure 4-4 plots the proficiency distributions in the requirement construct and the target construct based on the change-point SCM for the example given in Figure 4-2 (b). There is a single normal distribution in the requirement construct, and there are three distributions in the target construct for each group of persons on Level 1 (black line), Level 2 (red line), and Level 3 (blue line). As formulated in Equation (1.7), the examinees are divided into three groups based on the estimated cut scores $\gamma_1$ and $\gamma_2$. In particular, Level 2 examinees in the requirement construct are more likely to achieve Level 2 in the target construct, and Level 3 examinees in the requirement construct are more likely to reach Level 3 in the target construct. The possibility of different advantages for each group is shown by the two discontinuity parameters in Equation (1.8), $\lambda_1$ and $\lambda_2$, which each increase the probability of getting better scores on the items in the target construct. Similarly, according to the hypothesized links, the discontinuity parameters are likely to have positive values, although negative discontinuity parameters can occur, at least in the mathematical model. In addition, according to the constraints on the discontinuity parameters, the distribution of the target construct would show different patterns. As will be illustrated in the empirical example, if one link is not valid and the expected advantage does not exist ($\lambda_1=0$), only two distributions on the target construct, combination of the black line and the blue line, would be presented.
4.2.3. Estimation

A Markov chain Monte Carlo (MCMC) estimation was selected to estimate the parameters of the proposed change-point SCMs. WinBUGS 1.4.3 (Lunn et al., 2000) software can be used for this purpose. In order to implement the MCMC algorithm using WinBUGS, the prior distributions must be specified for all parameters, which include the cut score parameter, the discontinuity parameter, the item parameters, and the variance-covariance of the person proficiencies. Although each parameter can have different prior distributions, this study limits its scope to the use of simple and straightforward distributions, such as the conjugate prior distributions that make the posterior distribution belong to the same family. Following the conventions of Bayesian item response modeling (Curtis, 2010), a normal distribution was used for the fixed-effects parameters, including item parameters and discontinuity parameters. However, for the cut score parameter, a mildly informative prior, which is a uniform distribution was specified following Western and Kleykamp (2004). The range was set from -6 to 6 based on the typical range of item and person estimates on the logit scale. An inverse-Wishart distribution, which is a conjugate prior of the variance and covariance of the multivariate normal distribution, was specified for the variance and covariance matrix of the latent ability variables in the two dimensions (Gelman et al., 2013). For the model selection, the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) was used.

\[
\gamma_{(j)} : \text{Unif}(-6,6), \quad \lambda_{(j)} : N(0,1),
\]
\[
\beta_{rt} : N(0,1), \quad \tau_{rt} : N(0,1),
\]
\[
\tau_{I,R} : N(0,1), \quad \tau_{T,I} : N(0,1),
\]
\[
\Theta | \Sigma : \text{MVN}(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_{RT}^2 & \sigma_{RT}^2 \\ \sigma_{RT}^2 & \sigma_T^2 \end{bmatrix} : \text{Inverse-Wishart}(\Psi, \upsilon)
\]

4.3. Simulation Study

4.3.1. Data Generation

A simulation study was designed to evaluate the recovery of the parameters of the proposed change-point SCM. The Two Level Case data were generated using Equations (1.4) and (1.5), and the Three Level Case data were generated using Equations (1.7) and (1.8). For the data generation, the R software (R Development Core Team, 2013) was used. As illustrated in the model formulations, two constructs, called the requirement and target constructs, were specified, and examinees were classified into one of the two levels or the three levels based on the fixed cut score(s) in the requirement construct. The number of examinees was set to 1,000, and the latent abilities of the examinees in the two constructs
were assumed to follow a bivariate normal distribution: 

\[
\begin{bmatrix}
\theta_{pR} \\
\theta_{pT}
\end{bmatrix}
\sim MVN\left(
\begin{bmatrix}
0 & 0 \\
1 & 0.5
\end{bmatrix}
\right).
\]

The number of items in each construct was 30 (\(I_R = I_T = 30\)), and the parameters related to the items (i.e., item/step difficulties and deviations from the cut score(s)) were generated from a standard normal distribution. Specifically, for the Two Level Case, the mean of the item difficulties in the target construct was constrained to be zero (\(\beta_{IT} = -\sum_{i=1}^{29} \beta_{iT} \)), while there was no constraint for the deviations from the cut score in the requirement construct (\(\tau_{IR}\)). Similarly, for the Three Level Case, the mean of the item difficulties in the target construct (\(\beta_{IT} = -\sum_{i=1}^{29} \beta_{iT} \)) and the mean of the deviations from the cut scores in the requirement construct were constrained to be zero (\(\tau_{I1R} = -\sum_{i=1}^{29} \tau_{I1R} \)). In addition, for both constructs of the Three Level Case, the second step difficulty in the target construct was set to be negative of the first step difficulty (\(\tau_{I1R} = -\tau_{I2R}, \tau_{I1T} = -\tau_{I2T}\)) so that the step difficulty within each item were summed to be zero. More importantly, for the Two Level Case, the cut score parameter was specified as \(\gamma = 0\) and discontinuity parameter was specified as \(\lambda = 1\). The positive discontinuity parameter can be considered to indicate an advantage for examinees who were classified into the proficient level on the requirement construct. For the Three Level Case, the cut score parameters were specified as \(\gamma_1 = -0.5\) and \(\gamma_2 = 0.5\) with corresponding discontinuity parameters specified as \(\lambda_1 = \lambda_2 = 1\).

### 4.3.2. Analysis

After the data were generated, each dataset was analyzed using the proposed change-point SCM to assess the recovery of the parameter. For the estimation of the models, a Bayesian approach using an MCMC algorithm was implemented using WinBUGS 1.4.3 (Lunn et al., 2000). For all models, three chains with dispersed starting values were run, and the convergence of the chains was determined with the \(\hat{R}\) index with a critical value of 1.1 (Gelman & Rubin, 1992). In order to avoid high autocorrelations between the sampling distributions, the thinning interval was increased, and the posterior mean estimates were obtained from 25,000 iterations of post-burn-in after 25,000 iterations of burn-in. A total of 30 replicates were made for each Two Level Case and Three Level Case, and the bias and root mean square errors (RMSE) were calculated.

In addition, in an effort to validate the estimation of the cut score parameter, eight incorrect values of the cut score including the generating value (\(\gamma = -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0\)) were selected and fixed during the analysis instead of being estimated. For this analysis, one randomly chosen simulated dataset among 30 replicates was used. After fitting the change-point SCM with each fixed cut score, the model fit statistics using DIC and the model parameter estimates were compared.
4.3.3. Results

Recovery of the model parameters

Table 4-1 provides the recovery results of change-point SCM compared with the baseline models, which are RMRM for the Two Level Case and the RMPCM for the Three Level Case. For the item parameters, including the overall item difficulties in the target construct \( \beta_{iT} \), the deviations from the cut scores in the requirement construct \( \tau_{ir}, \tau_{i1R} \), and the item-specific step difficulties in the target construct \( \tau_{i1T} \), the table presents the averages and the standard deviations of the estimated bias and RMSE values of 60 items.

In general, the variance of the requirement construct \( \sigma^2_R \) and the overall item difficulties in the target construct \( \beta_{iT} \) were recovered well in both the baseline models (RMRM, RMPCM) and the change-point SCMs for the Two Level Case and the Three Level Case. The bias and the RMSE values of those parameters did not differ significantly; however, the change-point SCM that incorporated the discontinuity parameter yielded a much smaller bias and RMSE values with respect to other parameters, suggesting that the parameters appeared to be recovered better by the generating change-point SCMs.

In detail, for the Two Level Case, when the change-point SCM was applied, the bias ranged from -0.047 to 0.068, and the RMSE values ranged from 0.021 to 0.115. Even though it was small, among the model parameters, the largest negative bias was observed from the cut score \( \gamma \) and the discontinuity parameter \( \lambda \) as -0.047 and -0.011, respectively. For the Three Level Case change-point SCM, the bias ranged in magnitude from -0.034 to 0.041, and the RMSE values ranged from 0.014 to 0.178. Interestingly, the first discontinuity parameter \( \lambda_1 \) showed a small overestimation as 0.032, while the second discontinuity parameter \( \lambda_2 \) revealed a small underestimation with the bias value of -0.034. The opposite pattern was observed for the cut score parameter: the bias for the first cut score \( \gamma_1 \) was -0.002, while the bias for the second cut score \( \gamma_2 \) was 0.002. Overall, these results suggest that the proposed change-point SCM appeared to recover the generating values fairly well.

When the discontinuity parameters were ignored as in the baseline RMRM and RMPCM, the cut score parameter \( \gamma \) was considerably underestimated for the Two Level Case, but recovered quite well for the Three Level Case. This suggests that more examinees are likely to be classified in the proficient level, particularly for the Two Level Case, if the hypothesized link is not correctly specified in the model. More importantly, consistent with Choi’s (2013) findings, the impacts of neglecting discontinuity parameters were most obvious on the variance of the latent ability in the target construct \( \sigma^2_T \) and in the covariance between the two latent constructs \( \sigma_{RT} \). As shown in Table 4-1, the magnitude of the bias of \( \sigma^2_T \) were very large as 0.907 for the Two Level Case and 0.545 for the Three Level Case, suggesting that the variance of the target construct was considerably overestimated compared with the corresponding true value. Consequently, \( \sigma_{RT} \) was estimated to be much greater than the true value and the correlation between the two constructs \( \rho_{RT} \) was also overestimated. These results indicate that failing to incorporate the discontinuity parameters can be misleading in
terms of more variability in the target construct variance and the higher association between the latent variables in the two constructs than there actually are.

Table 4-1. Bias and RMSE in the baseline models and the change-point SCM

<table>
<thead>
<tr>
<th></th>
<th>Two Level Case</th>
<th></th>
<th>Three Level Case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
</tr>
<tr>
<td>$\sigma^2_R$</td>
<td>0.051</td>
<td>0.056</td>
<td>$\sigma^2_R$</td>
</tr>
<tr>
<td>$\sigma^2_T$</td>
<td>0.907</td>
<td>0.909</td>
<td>$\sigma^2_T$</td>
</tr>
<tr>
<td>$\sigma_{RT}$</td>
<td>0.540</td>
<td>0.540</td>
<td>$\sigma_{RT}$</td>
</tr>
<tr>
<td>$\rho_{RT}$</td>
<td>0.249</td>
<td>0.249</td>
<td>$\rho_{RT}$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-0.476</td>
<td>0.477</td>
<td>$\gamma_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\gamma_2$</td>
</tr>
<tr>
<td>$\beta_{iT}$ *</td>
<td>0.000</td>
<td>0.071</td>
<td>$\beta_{iT}$ *</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.008)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>$\tau_{ir}$ *</td>
<td>0.218</td>
<td>0.229</td>
<td>$\tau_{i1R}$ *</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.015)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\tau_{iT}$ *</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.037)</td>
</tr>
<tr>
<td>$\sigma^2_R$</td>
<td>0.005</td>
<td>0.024</td>
<td>$\sigma^2_R$</td>
</tr>
<tr>
<td>$\sigma^2_T$</td>
<td>0.013</td>
<td>0.043</td>
<td>$\sigma^2_T$</td>
</tr>
<tr>
<td>$\sigma_{RT}$</td>
<td>0.004</td>
<td>0.024</td>
<td>$\sigma_{RT}$</td>
</tr>
<tr>
<td>$\rho_{RT}$</td>
<td>0.000</td>
<td>0.021</td>
<td>$\rho_{RT}$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-0.011</td>
<td>0.042</td>
<td>$\lambda_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\lambda_2$</td>
</tr>
<tr>
<td>$\beta_{iT}$ *</td>
<td>0.000</td>
<td>0.071</td>
<td>$\beta_{iT}$ *</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.008)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>$\tau_{ir}$ *</td>
<td>0.042</td>
<td>0.087</td>
<td>$\tau_{i1R}$ *</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.015)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\tau_{iT}$ *</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.015)</td>
</tr>
</tbody>
</table>

* The averages of the bias and RMSE values of the estimated item parameters are presented, and the numbers in the parentheses correspond to the standard deviations of the bias and the RMSE.
Validation of the cut-scores

In an effort to validate the estimation of the cut score parameter, eight incorrect values of the cut score including the generating value (γ = -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0) were selected and fixed during the analysis for one randomly chosen simulated dataset. There was a clearly decreasing pattern in the DIC values as the cut score moves closer to the generating value, which is zero, and the change-point SCM (indicated by est) showed the lowest DIC value. Figure 4-5 presents the resulting estimated MCMC model parameter estimates for the Two Level Case. Dotted lines indicate the generating values used for the particular data set. For comparison purposes, the empty dot labeled est, which is the result from the change-point SCM for the same data set, is presented on the cut score axis in the same figure.

For the discontinuity parameter of the Two Level Case, the true value was set to 1.0. As expected, the change-point SCM and the cut score of zero yielded the closest estimates to the correct value, while the discontinuity parameter estimate tended to be underestimated when the incorrect fixed cut score was far from the generating value.

A similar pattern is observed in the model parameters of the bivariate proficiency distribution. First, for this particular simulated data set, the variance of the requirement construct distribution was 0.97, which is closest to the results from the change-point SCM, and when the cut score was fixed as zero or -0.5. The variance of the requirement construct tended to be underestimated when the cut score was set smaller than zero, while it was overestimated when the cut score was set larger than zero. For the variance of the target construct distribution, estimates from the change-point SCM and the cut score of zero or 0.5 showed the nearest result to the generating value, while the corresponding estimates were overestimated under the incorrect cut scores. The estimated correlation between two constructs from the change-point SCM was the nearest to the generating value, while the correlation under incorrect cut scores tended to be overestimated.

For the Three Level Case, an identical pattern is observed as illustrated in Figure 4-6. Again, dotted lines indicate the generating values used for this particular data set. One difference from the Two Level Case is that there were two cut scores for the Three Level Case, and the generating values were -0.5 for the first cut score and 0.5 for the second score. Because the second cut score was constrained to be the negative of the first cut score, the generating value used in this figure is the one for the first cut score (-0.5). Among the various incorrect cut score values, one interesting exception is when the cut score was fixed to zero, which leads to separating examinees into virtually two groups instead of three groups. In this case, at the cut score of zero, there is a single discontinuity parameter instead of two discontinuity parameters. Similar to the Two Level Case, the estimated DIC value was the smallest for the change-point SCM and when the cut score was fixed to the generating value, which was -0.5.

Regarding the discontinuity parameter of the Three Level Case, the dots represent the discontinuity parameter at the first cut score (λ₁), and the triangles represent the discontinuity parameter at the second cut score (λ₂). The estimates from the change-point SCM and fixing
the cut score to the correct value provided similar results, which are the closest to the generating value, 1. It is clear that the discontinuity parameter was underestimated when the cut score was set to incorrect values, and it appears to be severely underestimated when the first cut score was fixed as positive values, which implies, for example, two peculiar cut scores of 0.5 for the first transition and -0.5 for the second transition.

Next, there were also radical differences in the variance and covariance of the person proficiencies for the Three Level Case. For the requirement construct, the variance tended to be overestimated when the first cut score was set to be negative (smaller than the generating value, -0.5), while the variance tended to be underestimated when the first cut score was set to be positive (larger than the generating value, 0.5); however, the variance in the target construct was likely to always be overestimated under the incorrect cut scores. These patterns in the variance estimates resulted in an overestimated covariance and correlation between the two constructs under the incorrect cut scores. The overestimation became much worse when the first cut score (generating value = -0.5) was set to be positive.

Overall, for both the Two Level Case and the Three Level Case, the results of the validation of the cut scores showed that the results from the change-point SCM were very close to when the cut score was fixed to the generating value. When the fixed cut score became further away from the generating value, the model fit became worse, and the model parameters appeared to be more biased.
Figure 4-5. Estimated model parameters for each cut-score (Two level Case)
4.4. Empirical Examples

4.4.1. Data Sources

Two empirical data sets were used for illustration of the use of the change-point SCMs. First, as illustrated in Figure 4-1, a subset of the ADM project data, which consisted of responses to the items of the two selected constructs, the Concept of Statistics (CoS) construct,
and the Chance (Cha) construct, was used. The ADM project data was analyzed in two ways: first for the Two Level Case ADM using the dichotomized responses (proficient versus non-proficient) and second for the Three Level Case using the original polytomous items (Level 1, Level 2, and Level 3)\(^1\). For the Three Level Case ADM, being on Level 3 in the CoS construct was considered to be the requirement to reach Level 3 in the Cha construct, as shown in the first link in Figure 4-1. This corresponds to the first link only between Level 3 in the requirement construct and Level 3 in the target construct in Figure 4-2 (b). For this data, 14 polytomous items \((j=1,2,3)\) for the CoS construct and 17 polytomous items \((j=1,2,3)\) for the Cha construct were included. For the Two Level Case ADM, the original polytomous responses on the same items were dichotomized, and being proficient in the CoS construct was considered to be the requirement to reach the proficient level in the Cha construct, as shown in Figure 4-2 (a). In both cases, due to the complex test form design, the number of items administered to each student was not the same. In both data sets, the responses from 489 middle school students who answered at least two items in each construct were analyzed.

The second empirical dataset came from a subset of the MSP Carbon project data (Jin, Shin, Johnson, Kim, & Anderson, 2015). This data measured teacher’s content knowledge (CK) and pedagogical content knowledge (PCK), which is referred to as the CK-PCK data. The CK construct describes how teachers develop their subject matter knowledge in science ranging from force-dynamic reasoning to matter and energy transformation. In relation to the CK construct, teachers are expected to be able to transform subject matter knowledge for the purpose of teaching, which is connected to both science content knowledge and general pedagogy knowledge. This PCK construct ranges from the content general approaches to targeting the transition from naïve ideas to scientific big ideas. In each construct, six polytomous items \((j=1,2,3)\) were used to measure the CK and PCK construct, respectively. The responses from 192 teachers who answered at least two items in each construct were analyzed. In the learning progression that describes the relationship between CK and PCK, it was assumed that teachers must achieve Level 3 in the CK construct in order to achieve Level 3 in the PCK construct. They also must reach Level 2 in the CK construct, in theory, in order to reach Level 2 in the PCK construct. This corresponds with the two hypothesized links in Figure 4-2 (b).

4.4.2. Analysis

In order to demonstrate the use of the proposed change-point SCM, two ADM data sets, Two Level ADM and Three Level ADM, and the Three Level CK-PCK were analyzed. For each dataset, the proposed change-point SCM and baseline models, which were the RMRM for the Two Level Case or RMPCM for the Three Level Case were analyzed. Note again that these baseline models correspond to a constrained version of the change-point SCM in which only the cut score parameter(s) were estimated and the discontinuity parameter(s) were constrained to be zero. Thus, comparing the proposed change-point SCM with the

---

\(^1\) Due to the very small number of the observed responses at the higher levels, the higher score categories were aggregated (CoS4 → CoS3 and Cha 4,5,6 → Cha3) resulting in three valid categories.
baseline models (i.e., RMRM or RMPCM) allowed us to investigate the consequences of ignoring the discontinuity parameters. In particular, for the Two Level Case, the change-point SCM in Equations (1.4) and (1.5), which incorporated one cut score parameter and one discontinuity parameter, can be compared with the models formulated in Choi (2013). Thus, we also employed Choi (2013)’s models to investigate the effect of estimating the cut score and the different way of parameterization.

For the Three Level Case, four types of change-point SCMs with different constraints on the hypothesized links were analyzed. The first change-point SCM analysis (SCM 1) incorporated two cut score parameters and two discontinuity parameters as described in Equations (1.7) and (1.8). While SCM 1 is the most generalized version and estimates two discontinuity parameters freely, the second change-point SCM analysis (SCM 2) additionally constrained that the discontinuity parameters at the two cut-scores were equivalent (λ₁ = λ₂). Two other restrictions on the discontinuity parameters were also investigated. The third change-point SCM analysis (SCM 3) constrained that the hypothesized link only existed between Level 2 in the requirement construct and Level 2 in the target construct without a link at Level 3 (λ₂=0). In contrast, the fourth change-point SCM analysis (SCM 4) constrained that the hypothesized link only existed between Level 3 in the requirement construct and Level 3 in the target construct without the link at Level 2 (λ₁ = 0).

The assumptions underlying the four different change-point SCMs cover a variety of relationships between the requirement construct and the target construct for the typical Three Level Case in the learning progression. Comparing these four different models will enable us to examine which assumption can be empirically verified for the used dataset in terms of the model fits. For example, the hypothesized link in the Three Level Case ADM corresponds to the SCM 4, because it does not specify the relationship at the Level 2. On the contrary, the hypothesized links in the Three Level Case CK-PCK corresponds with the SCM 1 with two hypothesized links at Level 2 and Level 3 respectively. Thus, in theory, it is expected that the SCM 4 for the Three Level Case ADM and the SCM 1 for the CK-PCK data would show the best model fit among the four different analyses. This comparison can serve as validity evidences to support or disprove the theoretical hypothesized links in the learning progression.

Similar to the simulation study, all models were estimated using WinBUGS with MCMC estimation. For the baseline models, the RMRM for the Two Level Case and the RMPCM for the Three Level Case, had the same priors except that the discontinuity parameters were excluded.

4.4.3. Results

Two Level Case for the ADM

For the Two Level ADM, the parameter estimates of the RMRM and the change-point SCM analysis are listed in Table 4-2. We fitted Choi (2013)’s models for comparison and the results are also listed in the same table. The worst model fit came from the RMRM, which estimated only the cut score and ignored the discontinuity parameter. Choi’s two models fixed
the cut score as zero and estimated only the discontinuity parameters. In terms of the estimated DIC, the change-point SCM shows the best model fit among the four different model formulations. This implies that the estimation of both the cut score and the discontinuity parameter is favored in this empirical data. Because the cut score estimate from the change-point SCM is not statistically different from zero and Choi’s models showed better fit than the RMRM, incorporating the discontinuity parameter than the cut score parameter appears more crucial for this empirical data.

Table 4-2. Parameter estimates and standard errors for the Two Level Case ADM

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Choi (2013)’s models ($\gamma=0$)</th>
<th>Estimating cut score</th>
<th>RMRM</th>
<th>Change-point SCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ</td>
<td>Est. SE</td>
<td>Est. SE</td>
<td>Est. SE</td>
<td>Est. SE</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>1.26 0.41</td>
<td>1.49 0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>0.66 0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_{\text{CoS}}^2$</td>
<td>1.70 0.25</td>
<td>1.75 0.24</td>
<td>1.86 0.26</td>
<td>1.72 0.24</td>
</tr>
<tr>
<td>$\sigma_{\text{Cha}}^2$</td>
<td>1.61 0.57</td>
<td>1.09 0.23</td>
<td>2.94 0.32</td>
<td>1.01 0.18</td>
</tr>
<tr>
<td>$\sigma_{\text{CoS,Cha}}$</td>
<td>0.85 0.44</td>
<td>0.32 0.23</td>
<td>1.55 0.21</td>
<td>0.25 0.16</td>
</tr>
<tr>
<td>$\rho_{\text{CoS,Cha}}$</td>
<td>0.51 0.23</td>
<td></td>
<td>0.66 0.19</td>
<td></td>
</tr>
<tr>
<td>Level1</td>
<td>48.0% 47.0%</td>
<td></td>
<td>33.5% 39.7%</td>
<td></td>
</tr>
<tr>
<td>Level2</td>
<td>52.0% 53.0%</td>
<td></td>
<td>66.5% 60.3%</td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>6533.21 6458.97</td>
<td>6596.33 6451.31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the RMRM, the cut score was estimated to be -0.05 with a standard error of 0.27. Although the cut score is not statistically different from zero, it resulted in 33.5% of examinees being in the non-proficient level and 66.5% of examinees being in the proficient level in the requirement construct. In the change-point SCM in which a discontinuity parameter was incorporated, the cut score ($\gamma$) was estimated to be -0.20. This cut score is lower than the RMRM, but resulted in 39.7% of examinees being in the non-proficient level and 60.3% of examinees being in the proficient level in the requirement construct. On the contrary, the cut score was constrained to be zero in Choi’s two models and classified approximately half the examinees to the non-proficient level and proficient level, respectively. The estimated negative cut score from the change-point SCM suggests that the items were somewhat easy overall for the examinees, and the classification of the proficient level and the non-proficient level needs to be determined at a lower logit than zero, although a 95% credible interval for the cut score included zero.

In the change-point SCM, the discontinuity parameter, $\lambda$, was estimated to be 1.63 with a standard error of 0.18. The positive value of the estimated discontinuity parameter in
particular suggests that the examinees on the proficient level of CoS had an advantage in reaching the proficient level in Cha as much as 1.63 logit on average. In comparison, the SCM 1 used in Choi (2013) in which two discontinuity parameters were incorporated for both the proficient level and the non-proficient level, the examinees on the proficient level of CoS had an advantage in reaching the proficient level in Cha as much as 1.63 logit on average. In comparison, the SCM used in Choi (2013) in which two discontinuity parameters were incorporated for both the proficient level and the non-proficient level, the examinees on the proficient level of CoS had a greater advantage in Cha ($\lambda_1=1.26$) compared with the examinees on the non-proficient level of CoS ($\lambda_2=0.66$) in terms of the magnitude of the discontinuity parameter. In particular, the advantage for the examinees on the non-proficient level of CoS ($\lambda_2$) was not significantly different from zero. This is reflected in the model fit comparison as well; when a single discontinuity parameter was incorporated, the model fit improved.

Regarding the item parameters, the estimates from the RMRM and the change-point SCM are illustrated in Figure 4-7 for ease of presentation. A comparison of the RMRM with the change-point SCM shows the effect of estimating the discontinuity parameter because the RMRM is a constrained version of the change-point SCM that assumes that the discontinuity parameter was zero. Although the cut score estimates were different between the RMRM and the change-point SCM, the item parameter estimates in both the requirement construct and the target construct appear to be quite consistent across models. Thus, as examined in the simulation study, we can conclude that ignoring the discontinuity parameter for the Two Level Case ADM had negligible impacts on the estimation of the item parameters.

![Graphs showing item parameter estimates](image)

**Figure 4-7. Comparison of the item parameter estimates using the Two Level ADM**

Finally, the bivariate latent person proficiency distribution estimated from the RMRM and change-point SCM is illustrated in Figure 4-8 with a solid line for RMRM and a dotted line for change-point SCM. The upper panel is for the requirement construct and the lower panel is for the target construct. In particular, the dotted curve in the lower panel combines the distribution of $\theta_{pT} + \lambda$ for examinees in the proficient level on the requirement construct. It is apparent from the graph that the latent ability distribution of the target construct was
increased by the discontinuity parameter and the examinees on the target construct are differentiated more clearly into two groups (i.e., show a bimodal distribution). Moreover, the variance of the latent ability in target construct ($\sigma_{\text{Cha}}^2$) and the covariance between the two constructs, ($\sigma_{\text{CoS,Cha}}$), were estimated to be smaller compared with the RMRM. As examined in the simulation study, this suggests that the RMRM has overestimated the variance of the target construct and incorporating the discontinuity parameter in the change-point SCM has modified it appropriately.

![Figure 4-8. Distribution of bivariate person proficiencies using the Two Level ADM](image)

Three Level Case for the ADM

For the Three Level ADM, the parameter estimates of the RMPCM and the four change-point SCM analyses are listed in Table 4-3. Overall, the estimated DIC indicates that the SCM 4, which specified the link only at the higher level between Level 3 of the CoS and Level 3 of the Cha and is aligned with the learning progression, fits better than RMPCM and all other types of SCMs with different constraints. This suggests that the hypothetical link in the ADM learning progression is supported empirically over the other restricted versions of hypothesized links.
The RMPCM and the different types of the SCM analyses estimate the cut scores, and the estimated cut scores appear to be quite consistent across the different SCMs. Specifically, for the RMPCM, the estimated cut score was not significantly different from zero and resulted in only 4.7% of examinees being on Level 1, 9.4% of examinees being on Level 2, and 85.9% of examinees being on Level 3 in the requirement construct. In contrast to the baseline RMPCM, for the all change-point SCM analyses, two cut scores were statistically different from zero. Compared with the RMPCM, four types of change-point SCMs resulted in larger proportions in Level 2 and smaller proportions in Level 3. In the SCM 4, the cut scores were estimated as -0.38 and 0.38, which shows the largest gap among the four different types of SCM analyses. Overall, when the incorrect constraints were imposed on the hypothesized links as exemplified in the SCM 1, SCM 2, and SCM 3, the estimated cut scores were shrunken toward zero compared with the best fitting SCM 4. Based on the cut scores estimated from the SCM 4, only 3.5% of examinees were classified as Level 1, 18.4% of examinees as Level 2, and 78.1% of examinees as Level 3 in the requirement construct.

### Table 4-3. Parameter estimates and standard errors for the data of the three levels of ADM

<table>
<thead>
<tr>
<th></th>
<th>RMPCM</th>
<th>Change-point SCM 1</th>
<th>Change-point SCM 2 ((\lambda_1 = \lambda_2))</th>
<th>Change-point SCM 3 ((\lambda_2 = 0))</th>
<th>Change-point SCM 4 ((\lambda_1 = 0))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>Est.</td>
<td>SE</td>
<td>Est.</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td>-0.24</td>
<td>0.36</td>
<td>-0.31</td>
<td>0.05</td>
<td>-0.31</td>
</tr>
<tr>
<td>(\gamma_2)</td>
<td>0.24</td>
<td>0.36</td>
<td>0.31</td>
<td>0.05</td>
<td>0.31</td>
</tr>
<tr>
<td>(\lambda_1)</td>
<td>1.59</td>
<td>0.31</td>
<td>1.44</td>
<td>0.14</td>
<td>1.74</td>
</tr>
<tr>
<td>(\lambda_2)</td>
<td>1.25</td>
<td>0.22</td>
<td>1.44</td>
<td>0.14</td>
<td>2.07</td>
</tr>
<tr>
<td>(\sigma^2_{\text{CoS}})</td>
<td>2.04</td>
<td>0.23</td>
<td>2.05</td>
<td>0.23</td>
<td>2.08</td>
</tr>
<tr>
<td>(\sigma^2_{\text{Cha}})</td>
<td>4.03</td>
<td>0.40</td>
<td>1.61</td>
<td>0.33</td>
<td>1.38</td>
</tr>
<tr>
<td>(\sigma_{\text{CoS,Cha}})</td>
<td>2.67</td>
<td>0.24</td>
<td>1.51</td>
<td>0.23</td>
<td>1.36</td>
</tr>
<tr>
<td>(\rho_{\text{CoS,Cha}})</td>
<td>0.93</td>
<td></td>
<td>0.83</td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td>Level1</td>
<td>4.7%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Level2</td>
<td>9.4%</td>
<td>19.8%</td>
<td>20.0%</td>
<td>18.0%</td>
<td>18.4%</td>
</tr>
<tr>
<td>Level3</td>
<td>85.9%</td>
<td>75.7%</td>
<td>75.5%</td>
<td>78.5%</td>
<td>78.1%</td>
</tr>
<tr>
<td>DIC</td>
<td>9196.05</td>
<td>9048.77</td>
<td>9058.18</td>
<td>9093.31</td>
<td>9024.97</td>
</tr>
</tbody>
</table>

More importantly, with the constraint imposed in the SCM 4, the mean advantage for the Level 3 examinees of the CoS to attain Level 3 in the Cha was larger \((\lambda_2 = 2.07)\) than any other discontinuity parameter estimates from the other types of SCMs. Although the SCM 4 was the best fitting model, other types of SCMs show interesting results when the incorrect
constraints were made on the hypothesized links. For the SCM 1 in which two different discontinuity parameters were incorporated, \( \lambda_1 \) and \( \lambda_2 \), they were estimated as 1.59 and 1.25, respectively. Both discontinuity parameters had positive values, implying that examinees on Level 2 and Level 3 of the Cos are likely to have an advantage in reaching Level 2 and Level 3 of the Cha. The comparison of the magnitude of the discontinuity parameters in particular reveals that examinees on Level 2 of the CoS have a greater advantage in reaching Level 2 of the Cha \( (\lambda_1=1.59) \) compared with the examinees on Level 3 of the CoS in reaching Level 3 of the Cha \( (\lambda_2=1.25) \) even though the difference between them is not statistically different. The SCM 2 assumed that the discontinuity parameters at different cut points were equivalent \( (\lambda_1=\lambda_2) \). In other words, the advantage that the Level 2 examinees of the CoS had for reaching Level 2 of the Cha is equivalent to the advantage that the Level 3 examinees of the CoS had for reaching Level 3 of the Cha. It appears that in terms of the magnitude, the estimated discontinuity under this restriction is a type of weighted average of two different discontinuity estimates in the SCM 1. Based on the estimated DIC, the equivalence assumption imposed on the two links as described in the SCM 2 does not fit as well as the SCM 1. Next, the SCM 3 was analyzed to examine the effects of the opposite link, which only includes the link at the lower level between Level 2 of the CoS and Level 2 of the Cha. Under the constraint of the SCM 3, the discontinuity parameter was estimated to be 1.74 for the examinees on Level 2 of the CoS. As expected, the SCM 3, which presumed the opposite link only in the learning progression, showed the worst fit among the four different types.

Regarding the item parameters, the estimates from the RMPCM and the SCM 4, the best fitting model, is compared in Figure 4-9. The comparison of the SCMs with the RMPCM shows the effect of estimating the discontinuity parameters. The RMPCM can be viewed as a constrained version of the change-point SCM that assumes the discontinuity parameters to be zero. As can be seen, the item parameter estimates appear to be generally consistent between two different models, implying trivial impacts on the item parameter estimates when the discontinuity was neglected.
Similar to the Two Level ADM and as expected from the simulation study, the estimated variance of the latent ability in the Cha ($\sigma_{\text{Cha}}^2$) and the covariance between the two constructs were decreased compared with the corresponding values from the RMPCM. Again, this suggests that the RMPCM has overestimated the variance of the target construct and incorporating the discontinuity parameter in the change-point SCM has modified it properly. Furthermore, the correlations between the two constructs are now estimated much larger compared with the corresponding values from the Two Level ADM (0.19 vs. 0.92). This suggests that dichotomizing the original response categories may have decreased the relationships between two constructs than they actually are. The bivariate latent person proficiency distribution estimated from the RMPCM and change-point SCM4 is illustrated in Figure 4-10 with a solid line for RMPCM and a dotted line for change-point SCM4. The upper panel is for the requirement construct and the lower panel is for the target construct. In particular, the dotted curve in the lower panel combines the distribution of $\theta_{pT} + \lambda_2$ for examinees in the Level 3 on the requirement construct. It is apparent from the graph that the latent ability distribution of the requirement construct was pretty similar. On the other hand, the distribution of the target construct was increased by the discontinuity parameter and the examinees on the target construct are differentiated more clearly. Since one of the two links is not valid and the expected advantage does not exist ($\lambda_1=0$), only two distinctive modes are presented on the target construct.
Figure 4-10. Distribution of bivariate person proficiencies using the Three Level ADM

**Three Level Case for the CK-PCK**

For the Three Level CK-PCK data, the parameter estimates from the RMPCM and the four change-point SCM analyses along with the estimated DIC values are listed in Table 4-4. For this data, the underlying hypothesized link corresponds to the SCM 1 by which two different links between each level of the requirement construct and the target construct exist; however, this relationship was not successfully validated in terms of the estimated DIC values. Unlike the expectation of cognitive scientists who developed the learning progression (Jin et al., 2015), the SCM 4, which specified the link only at the higher level between Level 3 of the CK and Level 3 of the PCK, fit better than the other types of SCMs. This implies that the hypothetical link between Level 2 of the CK-PCK learning progression is not empirically supported well for this data.
### Table 4-4. Parameter estimates and standard errors for the three levels CK-PCK

<table>
<thead>
<tr>
<th>Est.</th>
<th>SE</th>
<th>Est.</th>
<th>SE</th>
<th>Est.</th>
<th>SE</th>
<th>Est.</th>
<th>SE</th>
<th>Est.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_1 )</td>
<td>0.00</td>
<td>1.00</td>
<td>2.01</td>
<td>0.17</td>
<td>1.98</td>
<td>0.17</td>
<td>2.49</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>( \lambda_2 )</td>
<td>2.03</td>
<td>0.16</td>
<td>2.01</td>
<td>0.17</td>
<td>0.29</td>
<td>0.09</td>
<td>0.39</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>( \gamma_1 )</td>
<td>-0.01</td>
<td>0.09</td>
<td>0.26</td>
<td>0.09</td>
<td>-0.26</td>
<td>0.09</td>
<td>-0.29</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>0.01</td>
<td>0.09</td>
<td>-0.26</td>
<td>0.09</td>
<td>0.52</td>
<td>0.13</td>
<td>0.74</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>( \sigma^2_{\text{CK}} )</td>
<td>0.75</td>
<td>0.15</td>
<td>0.53</td>
<td>0.12</td>
<td>0.23</td>
<td>0.11</td>
<td>0.24</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>( \sigma^2_{\text{PCK}} )</td>
<td>2.27</td>
<td>0.40</td>
<td>0.22</td>
<td>0.11</td>
<td>0.14</td>
<td>0.10</td>
<td>0.16</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>( \sigma_{\text{CK, PCK}} )</td>
<td>0.82</td>
<td>0.16</td>
<td>0.12</td>
<td>0.08</td>
<td>0.14</td>
<td>0.10</td>
<td>0.16</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>( \rho_{\text{CK, PCK}} )</td>
<td>0.63</td>
<td>0.36</td>
<td>0.40</td>
<td>0.47</td>
<td>0.47</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level1</td>
<td>32.3%</td>
<td>17.7%</td>
<td>16.6%</td>
<td>16.7%</td>
<td>16.7%</td>
<td>11.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level2</td>
<td>1.0%</td>
<td>47.9%</td>
<td>49.0%</td>
<td>49.5%</td>
<td>49.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level3</td>
<td>66.7%</td>
<td>34.4%</td>
<td>34.4%</td>
<td>33.8%</td>
<td>39.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>3772.86</td>
<td>3726.45</td>
<td>3727.68</td>
<td>3734.79</td>
<td>3692.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the RMPCM, the cut scores were estimated to be -0.01 and 0.01, respectively, which are not significantly different from zero. This implies that there are virtually only two distinctive levels rather than three levels. Based on these cut scores, 32.3% of the examinees were classified as Level 1, only 1% of examinees as Level 2, and 66.7% of examinees as Level 3. As implied by the cut scores, the baseline RMPCM was not successful in separating the examinees into three groups. On the contrary, in the SCM 4, the cut scores were estimated as -0.39 and 0.39, respectively, and the largest gap among the four different types of SCM analyses. These cut scores resulted in 11.5% of examinees being on Level 1, 49.5% of examinees being on Level 2, and 39.0% of examinees being on Level 3 of the CK construct. Furthermore, it is interesting to note that for the rest of the three types of SCMs, the cut score estimates showed the opposite result: the second cut score was estimated to be lower than the first cut score, which does not make sense in our current case. For comparison, in the SCM 1 in which two different discontinuity parameters were incorporated, the first cut score \( (\gamma_1) \) was estimated to be 0.26, and hence the second cut score \( (\gamma_2) \) was constrained to -0.26. These estimated cut scores have reversed results from the initial expectations of cognitive scientists who hypothesized the links between the specific levels and are consistent for the SCM 2 and the SCM 3.

Under the constraint of the SCM 4, the mean advantage for the Level 3 examinees of the CK in attaining Level 3 of the PCK was larger \( (\lambda_2=2.49) \) than any other discontinuity parameter estimates from the other types of SCMs. In comparison, the two discontinuity
parameters in the SCM 1, \( \lambda_1 \) and \( \lambda_2 \), were estimated as 0.00 and 2.03, respectively. The estimate of the first discontinuity parameter implies that no discontinuity occurs at the first cut point. On the other hand, SCM 2 and SCM 3 yielded similar discontinuity parameter estimates as 2.01 and 1.98.

For ease of presentation, the item parameter estimates from the RMPCM and the SCM 4 is illustrated in Figure 4-11. As can be seen, the overall item parameter estimates appear to be consistent between two different models, although there are some discrepancies in the right panel for the deviation from the cut scores for each item. We suspect that the discrepancies may be due to the small number of the items administered in this data, which is only six for each construct. In a future study, the minimum number of items that are required for the reliable estimation of the item parameters in the change-point SCMs would be needed.

![Figure 4-11. Comparison of the item parameter estimates using the Three Level CK-PCK](image)

Like the examples above and as predicted from the simulation study, the estimated variance of the latent ability in target construct \( (\sigma^2_{\text{pck}}) \) and the covariance between the two constructs were smaller than the corresponding values of the baseline RMPCM. We see that the RMPCM has overestimated the variance of the target construct and that incorporating the discontinuity parameters in the change-point SCM has reduced the variance. The bivariate latent person proficiency distribution estimated from the RMPCM and change-point SCM4 is illustrated in Figure 4-12 with a solid line for RMPCM and a dotted line for change-point SCM4. The upper panel is for the requirement construct and the lower panel is for the target construct. In particular, the dotted curve in the lower panel combines the distribution of \( \theta_{pT} + \lambda_2 \) for examinees in the Level 3 on the requirement construct. It is apparent from the graph
that the latent ability distribution of the requirement construct was quite similar. On the other hand, the distribution of the target construct was increased by the discontinuity parameter and the examinees on the target construct are differentiated more clearly. Similar to the Three Level ADM, it looks bimodal because one of the two discontinuity parameters was zero.

![Distribution of bivariate person proficiencies](image)

**Figure 4-12.** Distribution of bivariate person proficiencies using the Three Level CK-PCK

### 4.5. Conclusion and Discussion

The primary goal of this study was to describe and propose one possible way to formulate a measurement model for complicated learning progressions using the SCM approach based on the change-point analysis. For this purpose, this chapter briefly reviews previous statistical models, including the SCM based on ordered latent class models, and the SCM that assumes that each construct is a continuous latent variable and that utilized the predetermined cut score. The SCM proposed in this chapter is based on change-point analysis, which can be useful in estimating the cut scores and the discontinuities at the cut scores.

The hypothesized links specifically represent the assumption that reaching a particular level of the requirement construct boosts the probability to attain a certain level on the target construct. In the proposed change-point SCM, the hypothesized relations between levels across multiple constructs are modeled by incorporating the cut score parameters and the
discontinuity parameters. In brief, the two parameters of primary interest are (a) the change point that corresponds to the cut scores in the requirement construct leading to the classification of the examinees according to the cut scores, and (b) the magnitude of abrupt change at the change-point, which corresponds to the discontinuity at the cut scores that incorporates the mean change in the latent ability distribution in the target construct following the hypothesized links. The model formulation in this study allowed us (a) to estimate the cut score and the discontinuity at the cut score, (b) to identify examinees’ proficiency in terms of the level of the construct as well as the latent continuum (i.e., logit), and (c) to empirically validate the hypothesized link embedded in the learning progression by comparing the model fits using both correct and incorrect constraints.

In this chapter, the change-point SCM for the Two Level Case in two constructs with a single connection and the change-point SCM for the Three Level Case in two constructs with one or two connections are examined via simulation study and empirical data analysis. Results from the simulation study indicate that the proposed change-point SCM appears to recover the parameters quite well. In addition, the estimates from the change-point SCM are very close to when the cut score is fixed to the generating value. When the cut score is incorrectly fixed and becomes further from the true value, the model fit tends to become worse, and the model parameters, including the discontinuity parameter and the variance-covariance of the multidimensional constructs in particular, appear to be more biased. In the empirical example using the ADM project data and the CK-PCK data, this study illustrates the use of the proposed change-point SCM for validating the hypothesized link in the learning progressions. For the Three Level Case specifically, by comparing four change-point SCMs with different constraints on the hypothetical link, we found validity evidence to support or disprove the initial theoretical hypothesized links in the learning progression. However, through the application of the change-point SCM to the empirical data sets, it is difficult to determine the pattern in the model parameters when the model incorporates incorrect constraints on the links (i.e., generated by SCM1 but analyzed with SCM2, SCM3, and SCM4) or the number of items that are required for reliable estimation of the change-point SCM. Thus, future simulation study for these issues will be helpful.

The change-point SCMs proposed in this paper involve the Two Level Case and the Three Level Case, and it is straightforward to extend it to additional level situations by increasing the number of cut scores and discontinuity parameters. One limitation of the current study is that the proposed models assume only one direction in the hypothesized links, which are initiated from the requirement construct and terminated at the target construct; however, a more complicated case is when the requirement construct for one link becomes the target construct for other links (see the second link in Figure 4-1 from Cha6 to CoS4). In this case, the SCMs proposed in this chapter would not be directly applicable because there is no longer a clear distinction between the requirement construct and the target construct. Since this chapter is a preliminary study to investigate the ideas for SCM formulations based on the change-point analysis, exploration with the SCMs based on other approaches may be helpful to formulate those more complicated links.
Another limitation of the current study is that the score categories match the number of levels in the learning progressions. For example, for the Two Level Case, responses on every item were scored dichotomously, and for the Three Level Case, responses on every item were scored as Level 1, Level 2, and Level 3. This situation is not unusual given that the SCM framework is very likely to be grounded on a validated theory, careful item design, and scoring rubric, following such as the Berkeley Evaluation and Assessment Research (BEAR) Assessment Systems (Wilson & Sloane, 2000). However, there can be more complicated situations in which the score categories for all items or for a subset of items are different from the number of levels in the learning progressions. Polytomous responses, for example, can be used for the Two Level Case or dichotomous items can be used for the Three Level Case. These extensions for more complicated cases in terms of the response types and the relations between the constructs may enhance the application of the SCM modeling as a measurement model for complex learning progressions.
Chapter 5. Summary and Conclusion

In this dissertation, I investigated modeling rater effects and complex learning progressions by extending and applying multilevel and multidimensional item response models, with a primary focus on (1) detecting rater effects in double-scored performance assessments, (2) monitoring human raters with automated scoring engine, and (3) developing measurement models for complicated learning progressions. This dissertation consists of three papers, Chapters 2, 3, and 4. I present brief summary and conclusion of each chapter below.

In Chapter 2, we introduced the Rater Trifactor Model (RTM) that extends bifactor model to incorporate a cross-classified structure (e.g., items and raters) in addition to the general dimension for the multiple ratings data. Specifically, four types of RTMs depending on relaxation of the homoskedasticity assumption on a cross-classified structure (HOHO, HOHE, HEHO, HEHE) were described and investigated. In order to reflect the complexities in the real world assessments, such as incompleteness and imbalance, simulation studies were designed and conducted. In particular, we evaluated the impacts of missingness rate and consequences of ignoring the rater-specific measurement variances on parameter recovery. The result of simulation studies indicated that both bias and RMSE values became larger as the missingness rate increased. Given that the missingness resulted in a weak linkage, allocation of the items and rater assignments appeared crucial factors that affected the recovery of the generating values under the larger missingness condition. However, imposing homogeneous variances on either items or raters or both (HOHO, HOHE, HEHO) resulted in similar magnitudes of bias and RMSE values compared with the true HEHE. This implied robustness to ignoring the rater-specific measurement variances. For the empirical example of the Carbon Cycle project data, we found that the RTM HOHO fitted better than any other model, which suggested that participating raters were not significantly different in their measurement variances although they were different in their severities.

In Chapter 3, in order to examine whether the scores from automated engine (AE) can be used to monitor and provide feedback to human raters instead of expensive scores from human experts (HE), we analyzed empirical rater training data using two different mixed-effects ordered probit models. Comparison between the two models revealed that it is necessary to relax the homoskedastic measurement error variances across the raters. Furthermore, in order to relate the model parameters to the rater effects of our interest, fixed-effects rater location estimates were used for rater severity/leniency, and rater-specific measurement error variances were used for rater accuracy/inaccuracy. We also computed the thresholds using the resulting estimates to depict rater centrality/extremity.

For the data used in this study, the result revealed that the AE depicts the HE quite similarly in terms of the rater severity, while being slightly different in terms of the rater accuracy and rater centrality. In particular, the AE labeled same raters as lenient and severe as the HE did. However, AE classified fewer raters as demonstrating inaccuracy and centrality.
During the analyses, we fixed the model parameters associated with the HE and the AE, and used them as the criteria, but the range of the parameter estimates were larger compared to those fixed values, implying that the criteria used in this study were quite stringent.

Lastly, in Chapter 4, we described and proposed one possible way to formulate a measurement model for complicated learning progressions with hypothesized links using the Structured Constructs Model (SCM) approach based on change-point analysis. The hypothesized links specifically represented the assumption that reaching a particular level of the requirement construct boosted the probability to attain a certain level on the target construct. In the proposed change-point SCM, the hypothesized relations between levels across multiple constructs were modeled by incorporating the cut score parameters and the discontinuity parameters. The change-point SCMs proposed in this paper involved the Two Level Case and the Three Level Case. The model formulation in this study allowed us (a) to estimate the cut score and the discontinuity at the cut score, (b) to identify examinees’ proficiency in terms of the level of the construct as well as the latent continuum (i.e., logit), and (c) to empirically validate the hypothesized link embedded in the learning progression by comparing the model fits using both correct and incorrect constraints.

Results from the simulation study indicated that the proposed change-point SCM appeared to recover the parameters quite well. In addition, the estimates from the change-point SCM were very close to when the cut score was fixed to the generating value. When the cut score was incorrectly fixed and became further from the true value, the model fit tended to become worse, and the model parameters, including the discontinuity parameter and the variance-covariance of the multidimensional constructs in particular, appeared to be more biased. In the empirical example using the ADM data and the CK-PCK data, we illustrated the use of the proposed change-point SCM for validating the hypothesized link in the learning progressions. For the Three Level Case specifically, by comparing different change-point SCMs with incorrect constraints on the hypothetical link, we found validity evidence to support or disprove the initial theoretical hypothesized links in the learning progression.


Framework of Reference for Languages: Learning, teaching, assessment. Strasbourg, France: Council of Europe/Language Policy Division.


