Title
Cross-classified Random Effects Modeling for Moderated Item Calibration: An Application to English Language Proficiency Assessment for Special Population

Permalink
https://escholarship.org/uc/item/5996127r

Author
Chung, Seungwon

Publication Date
2019

Peer reviewed|Thesis/dissertation
Cross-classified Random Effects Modeling for Moderated Item Calibration: An Application to English Language Proficiency Assessment for Special Population

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Education

by

Seung Won Chung

2019
ABSTRACT OF THE DISSERTATION

Cross-classified Random Effects Modeling for Moderated Item Calibration: An Application to English Language Proficiency Assessment for Special Population

by

Seung Won Chung
Doctor of Philosophy in Education
University of California, Los Angeles, 2019
Professor Li Cai, Chair

Test forms are often modified in order to accommodate various special populations or situations where administration of the original test forms is infeasible. While this practice aims to promote fairness for all, the goal can only be met if they are coupled with a systematic method for obtaining comparable scores across test forms. Numerous psychometric barriers stand in the way. For example, the limited sample size of the students taking modified test forms can prevent the usage of standard calibration and linking procedures. One particular area in which these issues are pronounced is the English language proficiency (ELP) assessment for students who are blind or have low vision, and consequently take modified Braille test forms developed under English Language Proficiency Assessment for the 21st Century (ELPA21).

To address these bottlenecks, this study proposes a method for moderated item calibration. A unified cross-classified random effects model that jointly utilizes item response data to the original test form and judgmental data provided by expert raters is developed to revise item parameters for scoring modified test forms. Estimation of this new model is performed using the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). The method is programmed in R (R Core Team, 2018) and its implementation strategies are discussed in detail. The proposed model is applied to
Braille test forms in ELPA21, and its performance is compared to the common practice. Simulation study validates the modeling framework and provide guidance for future data collection and study design. More generally, this study is significant because of its broad adaptability to 1) any special population for whom direct item calibration or standard linking is not feasible, and 2) any operational or research setting where field testing cannot be conducted because of resource/sample size constraints.
The dissertation of Seung Won Chung is approved.

Mark P. Hansen

Noreen M. Webb

Michael H. Seltzer

Li Cai, Committee Chair

University of California, Los Angeles

2019
To Junsoo
CONTENTS

List of Figures ......................................................... ix

List of Tables .......................................................... x

Acknowledgments ..................................................... xi

Vita ................................................................. xiii

1 Introduction ......................................................... 1
  1.1 Background .................................................... 4
    1.1.1 An Overview of English Language Proficiency Assessment under ELPA21 ................................... 4
    1.1.2 Psychometric Issues .................................... 6
  1.2 Research Goal ................................................ 12
  1.3 Research Significance ...................................... 13

2 Cross-classified Random Effects Model ...................... 16
  2.1 Fixed Effects and Random Effects .......................... 16
  2.2 Cross-Classified Random Effects Model in Psychometrics ......................................................... 18
    2.2.1 Generalizability Theory ................................ 18
    2.2.2 Item Response Theory .................................. 19
    2.2.3 G theory and IRT ...................................... 22
  2.3 Cross-Classified Random Effects Model Under Mixed Effects Modeling ........................................... 22
    2.3.1 Linear Mixed Effects Model ............................ 23
    2.3.2 Nonlinear Mixed Effects Model ....................... 24
2.3.3 Generalized Linear Mixed Effects Model

3 Cross-classified Random Effects Model for Moderated Item Calibration

3.1 Data Structure

3.2 Model Formulation

3.2.1 Part I

3.2.2 Part II

3.2.3 Part III

3.3 Likelihood Functions

3.3.1 Part I

3.3.2 Part II

3.3.3 Part III

3.3.4 The Joint Likelihood

4 Estimation and Scoring for Cross-classified Random Effects Model

4.1 Model Estimation via the MH-RM Algorithm

4.1.1 Implementation of the MH Sampler

4.1.2 Complete Data Models and Derivatives

4.2 Estimation of Standard Errors

4.2.1 Recursively Approximated Standard Errors

4.2.2 Post-convergence Standard Errors

4.3 Scoring of Random Effects by Empirical Bayes

4.3.1 Recursively Approximated Random Effects Scores

4.3.2 Post-convergence Random Effects Scores

4.4 Additional Notes on Estimation Strategies

vii
5 Application to English Language Proficiency Assessment

5.1 Data and Methods

5.2 Results

6 Simulation Study

6.1 Simulation Study Design

6.2 Simulation Study Results

7 Discussions

7.1 Summary

7.2 Future Directions

Bibliography
### LIST OF FIGURES

1.1 A New Linking Methodology ................................................. 12

5.1 Observed Probability vs. Expected Probability: Listening ...................... 73
5.2 Observed Probability vs. Expected Probability: Reading ......................... 74
5.3 Item Characteristic Curves for Group 1: Grade 3 Reading (Items 1 - 8) ........ 75
5.4 Item Characteristic Curves for Group 1: Grade 3 Reading (Items 9 - 16) .... 76
5.5 Item Characteristic Curves for Group 1: Grade 3 Reading (Items 17 - 24) .. 77
5.6 Item Characteristic Curves for Group 2: Grade 3 Reading (Items 1 - 3) ...... 78
5.7 Item Characteristic Curves for Group 3: Grade 3 Reading (Items 1 - 8) ...... 79
5.8 Item Characteristic Curves for Group 3: Grade 3 Reading (Item 9) ............ 80
5.9 Item Characteristic Curves for Group 4: Grade 3 Reading (Items 1 - 2) .... 81

6.1 Accuracy of Standard Errors: Conditions 1 - 4 ................................... 94
6.2 Accuracy of Standard Errors: Conditions 5 - 8 ................................... 95
6.3 Accuracy of Standard Errors: Conditions 9 - 12 .................................. 96
6.4 Accuracy of Standard Errors: Conditions 13 - 16 .................................. 97
6.5 Accuracy of Standard Errors: Conditions 17 - 20 .................................. 98
6.6 Latent Ability Recovery ................................................................. 100
LIST OF TABLES

1.1 Performance Levels .................................................. 5
1.2 Overall Proficiency Levels ........................................... 6
1.3 Relationship between Braille Items and Online Items ............ 8
3.1 Data Structure for Original Calibration Data ....................... 28
3.2 Data Structure for Judgmental Data .................................. 29
5.1 Data Summary: Online Items .......................................... 59
5.2 Data Summary: Modified Braille Test Items ....................... 60
5.3 Item Summary for Cut 3: Probability of Correct Responses ........ 62
5.4 Item Summary for Cut 4: Probability of Correct Responses ........ 63
5.5 Item Summary: Relationship .......................................... 63
5.6 Rater Agreement: Listening ........................................... 65
5.7 Rater Agreement: Reading ............................................ 66
5.8 Parameter Estimates and Standard Errors ......................... 68
5.9 Scoring Table: Grade 3 Reading ...................................... 82
5.10 Scoring Table: Grade 3 Reading (Group 1 Only) .................. 83
6.1 Simulation Condition .................................................. 86
6.2 Estimates of Part I .................................................... 89
6.3 Estimates of Part II ................................................... 90
6.4 Estimates of Part III .................................................. 92
6.5 Recovery of Revised Item Parameters ............................... 99
ACKNOWLEDGMENTS

My first and foremost gratitude goes to my advisor and committee chair, Professor Li Cai. He has provided knowledge, insights, and resources throughout my Ph.D. years. Without his support, guidance, encouragement, and patience, this work could not have come to fruition. If I had to do a second Ph.D., I would do it with him again. I owe much to Professor Mark Hansen, who as ELPA21’s lead psychometrician, has kindly shared data sets and provided tremendous support for this dissertation. I am also thankful to the other members of my committee – Professors Michael Seltzer and Noreen Webb for their helpful questions and suggestions. I would be honored to be recognized worthy of Ph.D. by them. I would like to acknowledge Professor Stephen Sireci as well for his valuable comments on this work. I am also indebted to the ELPA21 team at CRESST.

I wish to express my gratitude to my former advisor, Professor Sun-Geun Baek, and other mentors who have opened the door to this field: Professors Hyun-Jung Park, Chanho Park, and Taehoon Kang. My gratitude extends to faculty members at and outside of UCLA who has given me intellectual and emotional support and helped my academic growth, including but not limited to Professors Minjeong Jeon, Mike Rose, José Felipe Martínez, Yingnian Wu, Nicolas Christou, Craig Enders, Jennifer Krull, Alberto Maydeu-Olivares, Peter Bentler, Dan McCaffrey, Guemin Lee, Tacksoo Shin, and Taehun Lee. In particular, I would like to recognize Professors Matthew Madison and Han Du for their help throughout my research.

In addition, I thank my former and current colleagues at the psychometrics lab at UCLA: Professors Ji Seung Yang, Scott Monroe, and Carl Falk; Drs. Moonsoo Lee, Megan Kuhfeld, Jane Li, Kat Schenke, and Emily Bo; and Larry Thomas, Sijia Huang, and Yonsoo Suh. Scott, Megan, Jane, Moonsoo, Larry, and Carl deserve special thanks for their help they have provided me during my baby years in Ph.D. Especially Scott was like a co-advisor to me in my first year. Kat also deserves special thanks for occasional
I want to thank my friends/colleagues I met at UCLA/CRESST – Professors Jongyeon Ee and Wes Bonifay; Drs. Nami Shin, Eunhee Keum, and Bo Pang; Jung Eun Yun, Jayashri Srinivasan, Priscilla Liu, Alana Kinarsky, Kristen Rohanna, Joy Zimmerman, Adrienne Dellinger, Eric Setoguchi, Meredith Langi, Deborah La Torre, Junok Kim, Shujin Zhong, and Zhaopeng Ding. My friends in Korea, Jin Ah Jung, Hyeyeon Kim, and Yoonmi Woo, also deserve my appreciation for their emotional support throughout my studies.

I wish to acknowledge generous financial support from Korea Foundation for Advanced Studies (KFAS) and the UCLA Department of Education. A travel grant from Society of Multivariate Experimental Psychology (SMEP) is also appreciated. The views expressed in the dissertation are my own and do not reflect the views or policies of these funding agencies.

Finally, to my parents, Ilhwan Chung and Chunbok Hwang, and lovely younger brother, Soochang Chung, had it not been for your endless support and sacrifice, I could not have come this far. My pursuit of education would not have happened if not for my dad, who has served in education for more than three decades and never stops trying to contribute to the betterment of Korean education. Lastly, to my husband and dearest Junsoo Park, how lucky I am to have the phenomenal person as my husband. Without him, I would never have been able to finish my dissertation and be able to begin my next academic journey. I wish I could find a better word than thank you. “May the stars shine upon the end of your road!” - *The Lord of the Rings.*
VITA

EDUCATION

2011 Bachelor of Arts, Education
Seoul National University, Korea

2013 Master of Arts, Education (Educational Measurement and Evaluation)
Seoul National University, Korea

WORK

2011–2012 Graduate Student Instructor
Seoul National University, Korea

2014–present Graduate Student Researcher
University of California, Los Angeles
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

2016–2017 Editorial Assistant
Journal of Educational Behavioral Statistics

PUBLICATIONS


CHAPTER 1

Introduction

The English learner (EL) population is continuously on the rise. In school year 2014-15, approximately 10% of all students attending public schools in the United States, or an estimated 4.6 million students, were ELs (McFarland et al., 2017). A considerable portion of the EL population are students with disabilities. Specifically, in 2014-15, students (aged 6-21) identified to have disabilities – meeting one of 13 federal categories of disabilities – comprised more than 10% of the EL population, or an estimated 560,000 (U.S. Department of Education, 2017; Wisconsin Center for Education Research, 2017).

Inclusion of ELs with disabilities in English language proficiency (ELP) assessment has long been federally mandated. Such laws are the Individuals with Disabilities Education Act (IDEA), the Elementary and Secondary Education Act of 1965 (ESEA), and Federal civil rights laws. For example, Titles I and III of the ESEA require states and local educational agencies (LEAs) to administer annual ELP assessment for all K-12 ELs (U.S. Department of Education, 2014). Recently, measures to properly accommodate this special population in education has become more concrete. In particular, the ESEA was amended by the Every Student Succeeds Act of 2015 (ESSA). New requirements include reporting the total number and percentage of ELs and those disaggregated at minimum by ELs with any disability (U.S. Department of Education, 2016). These recent legal adjustments reflect that developed societies must attend to this specific subgroups. Specifically, the society must establish assessment systems that meet the educational need of every single student with provision of equal opportunity. Egalitarianism, an educational ideology that a developed society ought to pursue and realize, is not only embodied in such federal laws but also invokes our responsibility to address the needs
of underprivileged learners.

In spite of the inclusive policy and the egalitarian ideology, fair administration of ELP assessment for ELs with disabilities is a challenging task in practice. In the testing context, equity and individual fairness translate largely to maintaining comparability across the student body (AERA, APA, & NCME, 2014; Camilli, 2006). Test fairness has been conceptualized in various ways. In essence, it is “a fundamental validity issue and requires attention throughout all stages of test development and use” (AERA et al., 2014, p.49). To that end, we must be able to respond to individual characteristics and ensure that interpretations of test scores are valid for the intended use. In other words, in order to uphold test validity in ELP assessment, we must respond to examinees’ disabilities such that 1) the conditions of testing are equitable, and 2) test scores carry the same meaning across ELs from the general population to ELs with disabilities.

The first point of ensuring equitable testing conditions resonates with concepts in testing such as accessibility, adaptation, accommodation and modification. Accessibility means that “all test takers should have an unobstructed opportunity to demonstrate their standing on the construct(s) being measured” (AERA et al., 2014, p.49). However, to improve the accessibility of subgroups, the design and/or administration of the original assessment must be altered. This refers to adaptation. Adaptation covers a variety of changes in, for instance, the content or presentation of test items, administration condition, and response processes. It can be differentiated in terms of comparability of scores: accommodation and modification. Accommodation refers to any variation in the assessment design or support with the aim of achieving more accurate measurement, which does not fundamentally change the test construct or affect the comparability of scores. On the other hand, modification fundamentally alters what the test measures and ultimately affects the comparability of scores (AERA et al., 2014). Sometimes accessibility and accommodation are not viewed as separate concepts but rather used closely together, however (Abedi & Ewers, 2013). The notion of modification suggests that the second point of the aforementioned validity issue, namely the comparability of test
scores between ELs from the general population and ELs with disabilities, may be un-
fortunately compromised.

Another important aspect of ELP assessment is the manner in which the test re-
results are used. According to the Survey of the States’ LEP (limited English proficiency) 
Students, 94% of those surveyed used the ELP test for EL classification and placement 
(Abedi, 2008). Proper classification is critical since inaccurate classification fails to afford 
appropriate education and may have negative impact on students’ learning experiences 
and outcomes. Clearly, a high-stakes decision hinges on ELP assessment, and hence 
reliable and valid ELP test scores must be provided. In methodological terms, the cut-
scores, the scores at which the levels of proficiency are distinguished, should be carefully 
determined such that fairness is ensured.

The difficulty of fair ELP assessment arises because, while disability must be pre-
vented from affecting score reporting and proficiency determination, ELs with disabili-
ties inevitably take modified test forms. Yet, the standard setting for proficiency levels is 
based on test forms taken by ELs from the general population and the item parameters 
from those tests. In addition, achievement-level descriptors are developed under the as-
sumption that sampling is from ELs from the general population. Therefore, the set of 
cut-scores employed for ELs from the general population cannot be applied to ELs with 
disabilities. As such, how to determine the cut-scores for ELs with disabilities is at the 
core of maintaining the comparability of scores of modified test items.

The context and the problems hereby presented are embodied in the English Lan-
guage Proficiency Assessment for the 21st Century (ELPA21). ELPA21 is a large-scale 
assessment system with the goal of measuring ELP for ELs by implementing newly 
adopted ELP standards (see Council of Chief State School Officers, 2014 for details). This 
research in fact stems from measurement issues in the conduct of the ELPA21 assessment

---

1Two consortia, Assessment Services Supporting ELs through Technology Systems (ASSETS) and En-
glish Language Proficiency Assessment for the 21st Century (ELPA21), have been developed for ELP 
assessments through the U.S. Department of Education’s Enhanced Assessment Grants (Guzman–Orth, 
Laitusis, Thurlow, & Christensen, 2016).
system. Accordingly in the following section, we provide background on ELPA21 and introduce a specific psychometric problem this study attempts to solve.

1.1 Background

1.1.1 An Overview of English Language Proficiency Assessment under ELPA21

The assessment consists of four domains (Listening, Reading, Speaking, and Writing) and six gradebands (K, 1, 2-3, 4-5, 6-8, and 9-12). There are two types of assessments: screener and summative. The screener assessment is used for placement of and instructional decisions for incoming ELs. The summative assessment is administered annually at the end of an academic year for continuation or exit decisions. Both types of assessments are administered online, for ELs from general population taking the original test forms.

Students’ domain scale scores are obtained from the ELPA21 primary scoring model, a four-dimensional correlated factor model with scale or item parameters obtained from item response theory (IRT). Detailed explanation on the multidimensional IRT (MIRT) modeling can be found in Cai and Hansen (2019). Using expected a posteriori (EAP) scoring, four posterior means are calculated. The scale scores in the four domains, i.e., domain scores, are then categorized into five performance levels per domain. The performance levels are: Level 1 (Beginning), Level 2 (Early Intermediate), Level 3 (Intermediate), Level 4 (Early Advanced), and Level 5 (Advanced). The description for each level is presented in Table 1.1.

The cut-scores and language proficiency levels were determined by conducting what is known as the standard setting procedure. The standard setting design specific to ELPA21 is the Bookmark method (see e.g., Lewis, Mitzel, & Green, 1999; Mitzel, Lewis, Patz, & Green, 2001). Panelist raters provided cut-score recommendations for every domain at two performance levels, Level 3 and Level 4, for six “grades” (K, 1, 3, 5, 7, and HS). This became the basis for the remaining performance levels and grades. For
Table 1.1: Performance Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: Beginning</td>
<td>Displays few grade-level English language skills and will benefit from EL Program support.</td>
</tr>
<tr>
<td>Level 2: Early Intermediate</td>
<td>Presents evidence of developing grade-level English language skills and will benefit from EL Program support.</td>
</tr>
<tr>
<td>Level 3: Intermediate</td>
<td>Applies some grade-level English language skills and will benefit from EL Program support.</td>
</tr>
<tr>
<td>Level 4: Early Advanced</td>
<td>Demonstrates English language skills required for engagement with grade-level academic content instruction at a level comparable to non-ELs.</td>
</tr>
<tr>
<td>Level 4: Advanced</td>
<td>Exhibits superior English language skills, as measured by ELPA21.</td>
</tr>
</tbody>
</table>

the five levels of proficiency, there are four cut-scores.

Based on the domain score combinations, the overall proficiency is categorized into three levels: Emerging, Progressing, and Proficient. If a student marks above or equal to Level 4 in all four domains, he/she is classified as ‘Proficient’; if a student marks below or equal to Level 2 in all four domains, he/she is classified as ‘Emerging’; ‘Progressing’ is assigned to a student with any other domain combination. Table 1.2 presents the description of each proficiency level.

In adherence to the inclusive policy, ELPA21 seeks to provide all ELs equitable access to assessment by addressing the needs of each individual. They provide two forms of accommodations, embedded and non-embedded accommodations, “for whom there is documentation of need on an IEP or 504 accommodation plan” (English Language Proficiency for the 21st Century, n.d. p.17). Embedded accommodations include unlimited recordings and replays. Non-embedded accommodations consist of assistive technology, large print paper test booklet, scribe, and speech-to-text, etc. (see the ELPA21 Accessibility and Accommodations Manual for details). In particular, ELPA21 has dedicated itself to
Table 1.2: Overall Proficiency Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emerging</td>
<td>Students are Emerging when they have not yet attained a level of English language skill necessary to produce, interpret, and collaborate on grade-level content-related academic tasks in English.</td>
</tr>
<tr>
<td>Progressing</td>
<td>Students are Progressing when, with support, they approach a level of English language skill necessary to produce, interpret, and collaborate on grade-level content-related academic tasks in English.</td>
</tr>
<tr>
<td>Proficient</td>
<td>Students are Proficient when they demonstrate a level of English language skill necessary to independently produce, interpret, collaborate on, and succeed in grade-level content-related academic tasks in English.</td>
</tr>
</tbody>
</table>

developing test forms for students who are blind or have low vision (BLV). It considers all types of visual impairments (e.g., total blindness, minimal light perception, restricted peripheral vision, varying degrees of low vision, etc.). The test form for those with BLV is basically Braille, accompanied by audio. For those items that are not easily converted to Braille, it developed an alternative format via the “twinning” process. An example would be replacing pictures with tangible three-dimensional objects. This study focuses on the Braille test form, which is available for students K through 12.

1.1.2 Psychometric Issues

Comparability of differing assessments across ELs from the general population to the BLV ELs are of concern. If we have different forms of tests that are intended to measure the same skills, we must make sure that all forms are on the same scale and lead to comparable reported scores. A pool of items originally developed for ELs from the general population are designed for online assessment. These items are adapted to Braille versions to test BLV students. The range of test adaptions is shown in Table 1.3. According to the classification of Braille items in Table 1.3, Group 1 would in general be considered to be product of accommodation. However, here we assume that all four
groups may lead to some difference in the meaning of scores. That is to say, all these changes herein are *modifications*. It must be mentioned that this assumption distinguishes our approach from the current standard procedure. In the standard approach, the same item parameters calibrated from the original assessment are used for Braille items in Group 1, ignoring any potential difference between the original assessment and the Braille form. This may be justified when Group 1 items are deemed invariant across different test form, that is, if they are deemed not to be *modifications*. This premise is questionable though, and consequently the common practice can be problematic if it is not true.

For the sake of test fairness, the psychometric properties of modified test items should be comparable to those of the original test items. When items are modified, however, they no longer retain the same psychometric properties. Specifically, the item parameters calibrated with responses on online assessment from the ELs from the general population would be inappropriate unless we make adjustments to the original item parameters. The relationship between the modified Braille test form and the original test form, and the resulting variation in the expected item difficulty with respect to the degree of modification are presented in Table 1.3. Naturally, we need to either separately calibrate the modified test form or apply a sort of linking process that promotes equity.² As shown in Table 1.3, the Braille version may not conform to the same content or statistical specifications.

²We use a generic term *linking* as opposed to *equating* because “equating adjusts for differences in difficulty, not differences in content” (Kolen & Brennan, 2004, p.3).
Table 1.3: Relationship between Braille Items and Online Items

<table>
<thead>
<tr>
<th>Group</th>
<th>Relationship</th>
<th>Item Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identical to the online item, other than necessary differences in the format of presentation; changes are superfluous to apprehending the meaning of the item.</td>
<td>Item difficulties should be close to the online item difficulty.</td>
</tr>
<tr>
<td>2</td>
<td>Similar to Group 1, but the braille form version response format is modified or a visual that orients students to the text has been deleted. For example, pictures that accompany a read-aloud story are dropped.</td>
<td>Item difficulties should be closely related to online, but may not be as a Group 1 items difficulty.</td>
</tr>
<tr>
<td>3</td>
<td>While these braille form items are based on the online item, revisions cause the depth of concept measured to be different from the online versions. For example, a graph in the online version is changed to a table.</td>
<td>The relationship to the difficulty of the online source item will vary, depending on the specific changes made.</td>
</tr>
<tr>
<td>4</td>
<td>These braille form items are designed to measure a concept from the ELPA21 standards and ALDs, but the items have different content from the online version. For example, an online item may ask students to compare two pictures, while the modified item measuring the same standards asks the student to compare two events.</td>
<td>There is no consistent relationship expected between item difficulties.</td>
</tr>
</tbody>
</table>

Unfortunately, neither the standard linking methods nor item calibration in the standard item response theory (IRT) is possible. For starters, a small sample size hinders item calibration. In the 2015-16 summative assessment, for example, a total 26 students, which amounts to only .007% of all students who participated in the assessment, took modified Braille test forms across all grades. Discussions of the sample size issue in the IRT literature suggest that only with a large sample size can the IRT item parameters be accurately estimated. This, in turn, improves the estimation of latent traits (DeMars, 2010). Although there is no “gold standard” or “magic number” regarding the exact sample size needed in IRT analyses (Morizot, Ainsworth, & Reise, 2007), and the sample size depends considerably on the choice of the IRT model, a large sample size is generally required.³ There appears to be a consensus that a number smaller than 100 is unacceptable even for the one-parameter logistic (1PL) model with dichotomous items for achieving stable parameter estimates (Embretson & Reise, 2000). Given that the number of BLV students is considerably smaller than 100, the item calibration is clearly not an option.

If the IRT item calibration is not feasible, then one may ask why we cannot apply a standard linking procedure instead. In the IRT scale linking procedure, two approaches are available: the common population approach and the common item approach (Yamamoto & Mazzeo, 2005). Obviously, common population linking is not appropriate because we can never have a same population of students that take both an original test form and a modified Braille test form. Common item linking is also inhibited due to the inherently small number of students who take modified Braille test forms even if we assume common items – that is, the item parameters of some items in the Braille version and the original version are treated equivalently. This is basically the same problem in the item calibration mentioned above. Importantly for this particular study, it should not be presumed that the item statistics of the Braille test forms would

³Note that other factors such as the study purpose (e.g., whether to evaluate questionnaire properties or to generate accurate IRT scores), the sampling distribution (e.g., how respondents are distributed over the range of construct), and the quality of items are to be considered as well (see Reeve & Fayers, 2005).
remain the same even when assessing the same tasks as the original test forms.

As a solution, ELPA21 has adopted a new linking methodology, namely “judgement-based, data-informed linking process” developed by Winter et al. (2018) for linking the two seemingly incompatible test forms. In Winter et al. (2018) under ELPA21, experts in education have gathered to provide the expected probabilities of correct responses to modified items from BLV students at cut 3 (differentiating Level 2 and Level 3) and cut 4 (differentiating Level 3 and Level 4). Recall that these two cuts are important because the overall proficiency level is immediately determined from them. The eventual goal is of course to overcome the comparability issues arising from basing the scores and classification of BLV students on the same set of criteria (cut-score, performance levels, and overall proficiency levels) for the ELs from the general population.

In the sense it measures judgment-based item statistics, Winter et al. (2018) is not entirely new per se. It evokes resemblance to some existing work, in particular anchor-based methods for judgmentally estimating item statistics (see Hambleton & Jirka, 2006). A more comprehensive literature review indicates that judgment-based approach is generally promising one (Farmer, 1928; Hambleton & Jirka, 2006; Lorge & Kruglov, 1952; Lorge & Diamond, 1954; Thorndike, 1982), and finally Hambleton and Jirka (2006) are all suggestive of the usefulness of involving judges (raters) to estimate (rate) item difficulty, even though it is not geared towards the goal of linking in those studies. Therefore, one may say that Winter et al. (2018) is the first to utilize the judgment-based approach for the purpose of linking. In turn, from the linking standpoint, we can find connections to “social moderation” (e.g., Linn, 1993; Mislevy, 1993).

Another resemblance can be drawn to standard setting. Notably, it is similar to Angoff standard-setting procedures but with a different purpose and a unique feature. The Angoff method, which was proposed by Angoff (1971), involves panelists with expertise who review each item and provide estimated proportions of correct responses from a population of interest, often the borderline examinees. A variation of the method goes through two rounds of the item judgment process, with discussions between the rounds.
The final cut-score is obtained by averaging all ratings, usually based on the final round (Cizek & Bunch, 2007).

While the Angoff method is a standard setting procedure to determine cut-scores, the linking methodology adopted by ELPA21 is to ultimately adjust item parameters for modified Braille test forms. Specifically, panelists provided expected probability of achieving an item score greater than or equal to a particular score point. For dichotomous items, this simply means the expected probability of correct responses. The procedure was performed over two rounds. In the first round, the judges provided independent ratings. In the second round, they revisited their scores to either keep their first-round ratings or change them after discussions among the judges. A unique part of the procedure is that, if the modified items have online counterparts, the panelists had been informed of the proportions of correct responses from the online assessment taken by ELs from the general population. Thus, this new method could be called “anchored Angoff method” (see also http://www.elpa21.org for details on this event). The approach compared to standard setting procedure used in online assessment can be visualized in Figure 1.1.

The next step is to estimate the item parameters for modified Braille test forms developed for BLV students from the panelists’ ratings collected via the cut-score linking procedure. While the methods in those utilizing raters’ judgements have taken linear fixed-effects approach (e.g., simple average of all the ratings) or pursued consensus among raters to obtain outcomes, either cut-score or any item statistics, this study attempt to propose an alternative way. This is where the motivation of this study stems from, and it is directly connected to the goal of this research.

---

4As mentioned earlier, ELPA21 used the Bookmark method for this purpose.
1.2 Research Goal

In this study, we seek a methodologically rigorous and defensible way to score special population, who take modified tests, when item calibration or standard linking is not feasible – a situation we have to commonly deal with in educational testing. To that end, this study aims to develop a unified cross-classified random effects model that utilizes item response data from the general population as well as judgmental data from expert raters in order to obtain revised item parameter estimates for use in scoring modified tests. We demonstrate this goal through an application to the context of Braille ELP assessment in ELPA21.

On the way to accomplishing the said main goal, this study introduces cross-classified random effects model in psychometrics under a comprehensive mixed effects modeling framework: a linear mixed effects model, generalized linear mixed effects model, and nonlinear mixed effects model. Furthermore, we develop an estimation strategy for the proposed unified cross-classified random effects modeling via Metropolis-Hastings Robins-Monroe algorithm (MH-RM; Cai, 2008, 2010a, 2010b). Finally, we investigate
the adequacy of the proposed modeling for *moderated item calibration* and the estimation strategy via a simulation study, in which conditions are varied in a manner reflective of the context of this study. We thereby provide some guidance for future data collection and design.

### 1.3 Research Significance

This research develops a unified model for moderated item calibration with random effects approach. This study contributes to the line of research on estimating item difficulty by judgmental methods (e.g., Farmer, 1928; Hambleton & Jirka, 2006) by providing an innovative methodological approach to moderated item calibration. The present literature and current practices in operation take the fixed effects approach (e.g., simple average of ratings from all expert raters) while continuously making enhancements in the judgmental process. This study, by comparison, enhances post-judgmental process by improving the statistical aspect of item calibration in two ways.

One is that the random effects approach taken in this study brings us significant benefits. We can “borrow strength” from other parts of data obtained from other items and raters. That is, with the random effects approach, even if we have highly unreliable ratings, our prediction for item parameters are not strongly affected because we are able to incorporate useful information from other data or namely *ancillary information* (Wainer et al., 2001). Therefore, random effects modeling (and the empirical Bayes estimation) will lead us to better estimates of item parameters. Another elaborate statistical treatment lies in the utilizing item response data from the general population in modeling. It allows the item parameters derived from raters’ judgments to be comparable to those of the original test items.

The moderated item calibration within our novel approach have wide applicability under varying practical and theoretical circumstances. One notable strength of moderated item calibration is that it is useful when standard linking is not feasible. Even
when we have common items or common population is assumed, our approach can certainly serve as an alternative path to the current small sample equating methods (see e.g., Albano, 2015; Kim, von Davier, & Haberman, 2008; Livingston & Kim, 2009; Skaggs, 2005). Previous studies have shown that when sample size is extremely small and the two test forms are not completely parallel, there is no sufficiently good solution for an equating method (Kolen & Brennan, 2004; Skaggs, 2005). It is also an alternative approach to setting new cut-scores for modified test forms for a special population in a classification setting, as is the context of this study. If one resorts to setting new cut-scores, the modified test scores cannot be interpreted in the same manner as the original test scores. By providing estimates of item parameters directly instead of setting new cut-scores, we enhance the comparability of scores. With the new item parameters, we are able to provide scores and profiles of proficiency on an equal footing across all population with the equivalent cut-scores of the original test forms. Consequently, it allows educators and administrators to properly classify special population by simply scoring their responses using the same criteria for students from the general population.

In addition, our approach may find extensive usage when resources are limited because our approach is potentially cost-saving. Determining the cut-scores require what is known as a standard setting procedure, in which raters must gather and discuss. This is a cumbersome and expensive ordeal. Likewise, collecting item properties from a judgmental method requires an equivalent amount of cost and time as the standard setting procedure. It would entail additional costs if more time is devoted to extensive training and feedback during the process. Our approach, however, does not require experts to gather and discuss for agreement in rating past a couple of rounds. We can further gauge the potential of this study’s approach in wider, non-educational contexts as well. A hypothetical situation would be where a drug company develops a drug for a rare disease that only a handful in the world have contracted and needs a labeling claim approval from the Food and Drug Administration (FDA). A direct calibration and field testing are infeasible in such a setting, and the findings of this study could offer a viable
This study expands the class of models to which the MH-RM algorithm can be creatively applied. The algorithm is already a proven computationally efficient estimation method in a number of research contexts (Cai, 2010a, 2010b; Falk & Cai, 2016; Monroe & Cai, 2014; Yang & Cai, 2014). Maximum likelihood estimation of generalized linear mixed effects model or nonlinear mixed effects models with crossed random effects used in this study have computational bottlenecks. Other existing methods are often not satisfactory in terms of accuracy if not efficiency. Overall, this study should establish a foundation for an alternative estimation method for cross-classified random effects model to other recent developments in this area (e.g., Cho & Rabe-Hesketh, 2011; Jeon, Rijmen, & Rabe-Hesketh, 2017). We note in passing that, because the focus of this study is on the accommodation of the MH-RM algorithm for a unified cross-classified random effects model for moderated item calibration, it does not examine the algorithm in-depth for each class of cross-classified random effects model.

Substantively, as this study is rooted in the pressing issue in the area of ELs with disabilities, its immediate consequences is that it allows computation of reliable and valid scores for modified Braille tests for BLV students in particular. In the end, this study contributes to the fruition of the ultimate purpose of the inclusive policy addressed in federal laws and regulations and the egalitarian ideology in education. This study addresses an important issue that our society must attend to, namely fairness in testing for all. In any assessment where different forms of testing are employed for students with disabilities or other underprivileged subgroups, the findings of this study will offer a level platform for evaluation.
CHAPTER 2

Cross-classified Random Effects Model

This chapter introduces cross-classified random effects model (CCREM), the foundation for model formulation in this study. This chapter has two major goals: 1) review of CCREM in general psychometric contexts; 2) establishment of CCREM within a mixed effects modeling framework. A mixed effects model is also known as a hierarchical or multi-level model.

The chapter begins with an overview of the concepts of fixed and random effects. Next, basic CCREMs in the measurement context, particularly within generalizability theory (G theory) and IRT, are presented. Lastly, conditional CCREMs are introduced. Specifically, linear mixed effects (LME) model, nonlinear mixed effects (NLME) model, and generalized linear mixed model (GLMM) (see e.g., Davidian & Giltinan, 1995; McCulloch, Searle, & Neuhaus, 2008; Vonesh & Chinchilli, 1997; Wu, 2010) will be discussed. While the notations are kept consistent as much as possible, in certain contexts some different notations may be introduced in order to comply with the conventions.

2.1 Fixed Effects and Random Effects

As a review of fixed effects and random effects, let us consider a simple one-way analysis of variance (ANOVA) fixed effects model and random effects model (Kirk, 1995). For \(i = 1, ..., I\) persons and \(j = 1, ..., J\) treatment levels, let \(Y_{ij}\) denote an observation for person \(i\) at treatment level \(j\). An experimental design model is written as

\[
Y_{ij} = \mu + \alpha_j + \epsilon_{ij},
\]  

(2.1)
where $\mu$ is the grand mean of the treatment-level population means and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the random error. The core difference between fixed-effect and random-effect approaches lies in how $\alpha_j$ is treated.

In a fixed effects model, $\alpha_j = \mu_j - \mu$ is the fixed treatment effect for population $j$, where $\mu_j$ is the population mean at treatment level $j$. Evidently, it is subject to $\sum_j \alpha_j = 0$, and $\mu$ is fixed as well, $\mu$ being constant across treatment levels and $\alpha_j$ constant for all observations at treatment level $j$. Importantly, a fixed effects model includes all treatment levels of interest. In a random effects model, $\alpha_j \sim N(0, \sigma^2_{\alpha})$ is a random variable indicating random treatment effect. A sample is considered random in two circumstances: 1) a sample is simply drawn randomly; 2) a researcher is willing to exchange conditions with any other sample of the same size from the population, provided that the sample size is much smaller than the size of the overall population (Shavelson & Webb, 1991).\(^1\)

Essentially, the decision to use which of the two approaches depends on whether or not treatment levels (or any factor levels) are to be considered random samples from a probability distribution (McCulloch et al., 2008). A random effects model is used when generalized conclusions must be drawn for more treatment levels than are included in the experiment (Kirk, 2014). From a statistical perspective, especially in estimation, an important advantage of random effects is that we are able to “borrow strength” from other groups or clusters in the prediction for an individual group or cluster. This point is connected to empirical Bayes (e.g., Raudenbush & Bryk, 2002; Thissen & Wainer, 2001). The benefits of using random effects approach come at a cost, of course. It requires stricter assumptions such as the distributional form assumption, constant variance, and zero correlations with included covariates.

It is worth noting for the sake of minimizing confusion that some literature avoid the terms “random effects” and “fixed effects,” instead opting for “varying” and “constant” (Gelman & Hill, 2007). In these terms, in fixed effects model, $\mu$ and $\alpha_j$ are constant over

---

\(^1\)The term “universe” is used instead of “population” in the original text of Shavelson and Webb (1991) and in G theory literature in general.
repeated samples and $\epsilon_{ij}$ varies over repeated samples. In random effects model, $\mu$ is constant over repeated samples, while $\alpha_j$ and $\epsilon_{ij}$ vary over repeated samples.

### 2.2 Cross-Classified Random Effects Model in Psychometrics

CCREM can be understood as a non-hierarchical model where units are cross-classified by more than one factor. As its name suggests, the model is formulated for the analysis of cross-classification designs. A prominent example of such a design is the typical data structure where students are classified both by the schools they attend and the neighborhood they live in (e.g., Goldstein, 1994; Raudenbush, 1993). In psychometrics, CCREMs arise within G theory and IRT. Traditional IRT models are basically mixed effects models with random person and fixed item, but if on top of it items are also treated as random, then they become crossed-classified random effects models. From this point onward, we only consider a two-way balanced cross-classification design. It is simple and also sufficient for the scope of this research.

#### 2.2.1 Generalizability Theory

"Generalizability theory is to measurement what the ANOVA is to substantive research (Shavelson & Webb, 1991, p.17).” As such, the analysis of a two-way cross-classification design in G theory is technically the same as the ANOVA with a two-way table. G theory is essentially a random facet measurement theory in the sense that it only considers designs with at least one random facet (Shavelson & Webb, 1991). In G theory, observed scores can be decomposed into effects using the random effects model or mixed effects models having a fixed facet. Here, CCREM is presented using the person-by-item ($p \times i$) design (Brennan, 2001; Shavelson & Webb, 1991).

For $p = 1, \ldots, P$ persons (or respondents) and $i = 1, \ldots, I$ items, let $Y_{pi}$ denote the

---

2In G theory, a facet refers to a major source of variation. In the person-by-item design, item is a facet of persons’ scores, and hence it is a one-facet measurement.
(observed) score for person \( p \) in the population on item \( i \) in the universe. First, we define a person’s universe score, \( \mu_p \), as

\[
\mu_p \equiv E_i Y_{pi},
\]  

(2.2)

where \( E_i \) is the expected value over items. Next, the population mean for an item, \( \mu_i \), is

\[
\mu_i \equiv E_p Y_{pi},
\]  

(2.3)

where \( E_p \) is the expected value over persons. Lastly, the grand mean over the population and universe is

\[
\mu \equiv E_p E_i Y_{pi}.
\]  

(2.4)

Having defined every component, we are now able to express the observed score using random effects model as follows:

\[
Y_{pi} = \mu + \nu_p + \nu_i + \nu_{pi}
\]  

(2.5)

where \( \nu_p = \mu_p - \mu \) is the person effect, \( \nu_i = \mu_i - \mu \) is the item effect and \( \nu_{pi} = Y_{pi} - \mu_p - \mu_i + \mu \) is the residual effect. The interaction effect and random error are confounded in the residual effect.

### 2.2.2 Item Response Theory

For the purpose of illustration, we shall introduce a unidimensional one parameter logistic (1PL) model for dichotomous responses. This is equivalent to the \( p \times i \) design with dichotomous responses, except that in the traditional IRT model, item is not considered as a source of variability. Again, we consider \( p = 1, \ldots, P \) persons and \( i = 1, \ldots, I \) items, now each scored dichotomously. Let \( y_{pi} \) denote the response of person \( p \) to item \( i \). Subsequently, we have two categories: \( y_{pi} = 1 \) if correct, and \( y_{pi} = 0 \) if not. Let \( \theta_p \) be the latent variable (e.g., latent ability) for person \( p \). With the conditional independence (Lord
& Novick, 1968), the probability of person $p$ responding correctly to item $i$ is written as

$$P(y_{pi} = 1|\theta_p) = \frac{1}{1 + \exp[-(\theta_p + b_i)]}, \quad (2.6)$$

where $-b_i$ is the item difficulty parameter for item $i$. Alternatively, the 1PL model can be formulated using logit link as

$$\text{logit}[P(y_{pi} = 1|\theta_p)] = \text{logit}(\pi_{pi}) = \theta_p + b_i, \quad (2.7)$$

where $\pi_{pi}$ is the probability that $y_{pi} = 1$, and $\theta_p \sim N(\mu_\theta, \sigma^2_\theta)$. For identification, it is a convention to define the mean of the distribution $\theta_p$ to be 0. The commonly used basic IRT model thus can be viewed as a fixed item, random person approach. To see how this model can be expressed in a hierarchical model, refer to Noortgate, De Boeck, and Meulders (2003).

Within the same $p \times i$ design, a fixed person, random item approach is also possible (e.g., Noortgate et al., 2003; De Boeck, 2008). In the 1PL case, item location (or the item difficulty parameter) is treated as random, and the person ability parameter is treated as fixed. De Boeck (2008) introduces random item effects model for three purposes: 1) measurement of people’s ability followed by a generalization of items, 2) explanation of item difficulties, and 3) differential item functioning (DIF). From the modeling perspective, the first point poses a direct contrast to the traditional model, while the second point extends the basic random item effects model. The third point suggests that random item effects model could have a variety of applications to psychometric problems.

Random item effects model is defined as

$$\text{logit}[P(y_{pi} = 1|b_i)] = \text{logit}(\pi_{pi}) = \theta_p + b_i, \quad (2.8)$$

where $b_i \sim N(\mu_b, \sigma^2_b)$. Notice that Equation 2.8 is analogous to Equation 2.6, except the roles of $\theta_p$ and $b_i$ are now exchanged; the response probabilities are conditional on $b_i$. 
not \(\theta_p\). Accordingly, we define the location distribution, not the ability distribution. As a side note, the notion of random item can be also found in a person-response curve (PRC; Trabin & Weiss, 1983), the approach of which comes from the psychometric studies of Lumsden (1977, 1978). A PRC reflects a person’s correct response probability to an item as a monotonically decreasing function of item difficulty (Reise, 2000).

Now let us define cross-classified random effects IRT model:

\[
\logit[P(y_{pi} = 1|\theta_p, b_i)] = \logit(\pi_{pi}) = \theta_p + b_i. \tag{2.9}
\]

Here, \(\theta_p \sim N(\mu_\theta, \sigma^2_\theta)\) and \(b_i \sim N(\mu_b, \sigma^2_b)\). Both person and item are regarded as random samples from populations. While this kind of a model is viewed as non-hierarchical in the sense that person and item are crossed at the same level, CCREM equation itself can be formulated either as a mixed effects model or hierarchical model (see Raudenbush & Bryk, 2002).

A further extension of cross-classified IRT model can be made through covariates. If the person parameter is treated as random, then the covariates that quantify the random person effect can be included in the model. A model with person covariates is in fact person explanatory IRT model (Wilson, De Boeck, & Carstensen, 2008). This is where a connection between IRT model and hierarchical linear model (HLM) originates (see e.g., Fox & Glas, 2001; Kamata, 2001; Kamata & Vaughn, 2011 for details on multilevel IRT model). By analogy, if the item parameter is treated as random, the covariates that quantify the item random effect may be incorporated in the model. This model may be called item explanatory IRT model. This relates to the second of the three aforementioned purposes for introducing random item effects model. This model stems from what is known as linear logistic test model (LLTM; Fischer, 1973). LLTM characterizes item difficulties as a linear function of item covariates. If an error term is added to LLTM, it becomes random item explanatory model.

Finally, person and item explanatory models can be combined to form cross-classified
doubly explanatory IRT model. This model will be discussed in detail under the mixed effects modeling framework in Section 2.3.3.³

2.2.3 G theory and IRT

Before moving onto the mixed effects modeling framework, it is worth briefly touching basis on these two theories. IRT and G theory have been regarded as incompatible testing theories. G theory is primarily a sampling model, whereas IRT is primarily a scaling model (Brennan, 2001; Briggs & Wilson, 2007). An important characteristic of G theory is the distinction between fixed and random effects, and that facets are flexibly designed as either fixed or random according to the study purpose. In standard IRT, items are treated as fixed while persons are treated as random. Accordingly, the most prominent difference between IRT and G theory in the conventional sense is that the items are fixed in IRT while they are usually treated as random in G theory, and hence the apparent incompatibility (Brennan, 2011). As detailed earlier, however, the current research is not limited to the random person, fixed item approach. There is an increasing trend towards utilizing cross-classified random effects IRT model. Furthermore, there has even been an attempt to formulate a direct correspondence between G theory and IRT by proposing IRT as a random effects model (Briggs & Wilson, 2007).

2.3 Cross-Classified Random Effects Model Under Mixed Effects Modeling

CCREMs with covariates, namely conditional CCREMs, fall under the umbrella of the mixed effects modeling framework. In this section, we introduce the general forms of CCREM in psychometrics through various types of mixed effects models (e.g., Davidian

³Wilson et al. (2008) describes explanatory models, but the explanatory features are only applied to the random person effects model, i.e., the traditional IRT model. Thus, terms such as item explanatory and doubly explanatory that we borrow from Wilson et al. (2008) in the following descriptions do not refer to the same models.
& Giltinan, 1995; McCulloch et al., 2008; Vonesh & Chinchilli, 1997; Wu, 2010). Limiting the scope to parametric models, we can formulate three different models depending on the type of response and the relationship between the response and cross-classified random effects: linear mixed effects model (LME), nonlinear mixed effects model (NLME), and generalized linear mixed effects model (GLMM). We shall consider only one observation in each cell of cross-classifications, which is a common structure in measurement. Extension to repeated measures within cell is possible and straightforward, and shall appear as moderated item calibration is modeled in Chapter 3.

2.3.1 Linear Mixed Effects Model

Continuing with the crossed \( p \times i \) design, let there be \( q = 1, ..., Q \) item properties and \( s = 1, ..., S \) person properties. Let \( X_1 \) and \( X_2 \) each denote the item and person covariates (with fixed effects). Specifically, let \( X_{1,iq} \) be the value for item \( i \) on item property \( q \) and \( X_{2,ps} \) be the value for person \( p \) on person property \( s \). For a single cell in the \( p \times i \) design, a linear CCREM with continuous outcomes incorporating the item and person covariates can be written as

\[
y_{pi} = \mu + \sum_{q=1}^{Q} \beta_{1,q} X_{1,iq} + \sum_{s=1}^{S} \beta_{2,s} X_{2,ps} + \zeta_i + \xi_p + e_{pi}, \quad (2.10)
\]

where \( \mu \) is the fixed intercept, \( \beta_{1,q} \) is the regression coefficient for \( X_{1,iq} \), \( \beta_{2,s} \) is the regression coefficient for \( X_{2,ps} \), and \( \zeta_i \) and \( \xi_p \) are the random residual effects of item and person after accounting for item and person covariates. It is assumed that \( \zeta_i \sim N(0, \sigma_{\zeta}^2) \) and \( \xi_p \sim N(0, \sigma_{\xi}^2) \). Finally, \( e_{pi} \sim N(0, \sigma_{e}^2) \) is the random error. Equation 2.10 can be written more compactly as

\[
y_{pi} = x_{pi}' \beta + z_{pi}' u_{pi}, \quad (2.11)
\]

where \( \beta = (\mu, \beta_{1}', \beta_{2}')' \), \( z_{pi} = (1, 1)' \), and \( u_{pi} = (\zeta_i, \xi_p)' \). Here, \( x_1' = (X_{1,i1}, ..., X_{1,iQ}) \) and \( x_2' = (X_{2,p1}, ..., X_{2,pS}) \) are a \( Q \times 1 \) vector of item covariates and a \( S \times 1 \) vector of person covariates, respectively, and \( \beta_1 = (\beta_{1,1}, ..., \beta_{1,Q})' \) and \( \beta_2 = (\beta_{2,1}, ..., \beta_{2,S})' \) are
$Q \times 1$ vectors of the regression coefficients for item covariates and an $S \times 1$ vector of the regression coefficients for person covariates, respectively.

Now we can write a matrix equation for the CCREM under LME. Recall that there are $P$ independent persons and $I$ items. Let $N = P \times I$ be the total number of responses, and $y = (y_{11}, ..., y_{PI})'$ denote an $N \times 1$ response vector by all persons on all items. The design matrix of the fixed effects is $X = [1_N|X_1|X_2]$, where $1_N$ is an $N$-dimensional unity vector, $X_1$ is an $N \times Q$ fixed item effects design matrix, and $X_2$ is an $N \times S$ fixed person effects design matrix. The design matrix of the random effects is $Z = [Z_\zeta|Z_\xi]$, where $Z_\zeta = 1_P \otimes I_I$ and $Z_\xi = I_P \otimes 1_I$, where $\otimes$ stands for Kronecker product. We also collect the two random effects in $u = (u_\zeta', u_\xi')'$, where $u_\zeta = (\xi_1, ..., \xi_I)'$ is an $I \times 1$ vector of the random item effects and $u_\xi = (\xi_1, ..., \xi_P)'$ is a $P \times 1$ vector of the random person effects. Let $e$ be an $N \times 1$ vector of the error terms. Using these notations, we recast the linear CCREM as an LME as follows:

$$y = X\beta + Zu + e. \quad (2.12)$$

We assume that $u$ has zero means and covariance matrix $\Psi$, and $e$ has zero means and covariance matrix $\Phi$. This leads to $E(y) = X\beta$ and $\text{var}(y) = ZZ' + \Phi$.

### 2.3.2 Nonlinear Mixed Effects Model

In order to accommodate the nonlinear relationship between response and random effects/covariates, we introduce NLME. Incorporating item and person covariates in the $p \times i$ design, we can express nonlinear CCREM in the following form:

$$y_{pi} = g(\phi_{pi}) + e_{pi}, \quad (2.13)$$

where $g(\cdot)$ denotes a nonlinear differentiable function, $\phi_{pi}$ is a linear function with cross-classified random effects, and $e_{pi}$ is the random error. The linear function $\phi_{pi}$ that maps
onto \( g(\cdot) \) is

\[
\phi_{pi} = \mu + \sum_{q=1}^{Q} \beta_{1,q} X_{1,iq} + \sum_{s=1}^{S} \beta_{2,s} X_{2,ps} + \xi_i + \xi_p,
\]

(2.14)

We can write Equation 2.15 more compactly as

\[
\phi_{pi} = x'_{pi} \beta + z'_{pi} u_{pi},
\]

(2.15)

Note that the linear function \( \phi_{pi} \) has the same form as Equation 2.11. The notations are not different from those in LME.

We can write the NLME in a matrix form. Again, let \( y = (y_{11}, ..., y_{PI})' \) denote an \( N \times 1 \) response vector by all persons on all items. We write the collection of the linear functions as \( \phi = (\phi_{11}, ..., \phi_{PI})' \) and its nonlinear-mapped function as \( g(\phi) = (g(\phi_{11}), ..., g(\phi_{PI}))' \). Then the nonlinear CCREM is recast as an NLME as

\[
y = g(\phi) + e,
\]

(2.16)

As in LME, \( u \) has zero means and covariance matrix \( \Psi \), and \( e \) has zero means and covariance matrix \( \Phi \). Note that response is assumed to be normally distributed (Wu, 2010).

### 2.3.3 Generalized Linear Mixed Effects Model

While in NLME the nonlinearity stems from the response not being linear in parameters, in GLMM a link function is what a model characterizes its nonlinear aspect (De Boeck & Wilson, 2004). The aforementioned doubly explanatory cross-classified random effects IRT model (see Section 2.2.2) is an example of GLMM. Let us introduce the doubly explanatory cross-classified random effects IRT model under the general formulation. Let \( y_{pi} \) be an independently distributed binary response given random effects \( u_{pi} \) from a distribution in the exponential family. Recall that \( u_{pi} = (\xi_i, \xi_p)' \). The conditional density
of $y_{pi}$ can be written in the exponential family form as follows:

$$f(y_{pi}|u_{pi}) = \exp \left\{ \frac{y_{pi} \gamma_{pi} - b(\gamma_{pi})}{a(\phi)} + c(y_{pi}, \phi) \right\}$$

(2.17)

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are known functions, $\gamma_{pi}$ is the location parameter, and $\phi$ is the scale parameter. The response would be independent Bernoulli, and thus $b(\gamma_{pi}) = \log[1 + \exp(\gamma_{pi})]$, $a(\phi) = 1$, and $c(y_{pi}, \phi) = 0$.

Once the distribution of response is specified, we relate the conditional mean response to the linear predictors (covariates) via some link function $g(\cdot)$. As worthy reminder, cross-classified random effects IRT model uses a logit link. Accordingly, we transform the conditional mean of $y_{pi}$ as $\mu_{pi} = E(y_{pi}|u_{pi}) = P(y_{pi} = 1|u_{pi})$, and write a linear model in the covariates as follows:

$$g(\mu_{pi}) = \mu + Q \sum_{q=1}^{Q} \beta_{1,q} X_{1,iq} + S \sum_{s=1}^{S} \beta_{2,s} X_{2,ps} + \zeta_{i} + \xi_{p}.$$  

(2.18)

where $g(\mu_{pi}) = \text{logit}(\mu_{pi})$. More compactly,

$$g(\mu_{pi}) = x'_{pi} \beta + z'_{pi} u_{pi}.$$  

(2.19)

Again, $g(\mu_{pi})$ has the same form as Equation 2.11 with the same notations defined in LME. As a side note, the two crossed random effects can be expressed in terms of two separate vectors such that $u_{p}$ is the person random effects and $v_{i}$ is the item random effects, for example (see Jeon et al., 2017).
CHAPTER 3

Cross-classified Random Effects Model for Moderated Item Calibration

Now that the generalities of CCREM has been delineated, we narrow its context to the scope of the study. Specifically, model formulation for moderated item calibration for modified (Braille) test forms is presented, followed by data structure we aim to model. Recall the goal of this study is to develop a unified CCREM that model judgmental data on modified (Braille) test items and original item calibrations together. As a quick overview, we propose a three-part model: 1) A traditional fixed item IRT model on responses on unmodified (original) items, 2) a cross-classified random effects IRT model on responses on modified (original) items, and 3) a multivariate nonlinear CCREM on judgements by raters.

3.1 Data Structure

Our data structure consists of two parts: the original calibration data (from the general EL population’s responses to the original test items) and the judgmental data for modified Braille test items (from a panel of expert raters). Without loss of generality, we only consider dichotomous responses/ratings. First up is cross-classification design of the original calibration data. Suppose there are \( p = 1, \ldots, P \) independent respondents, \( j = 1, \ldots, J \) unmodified items, and \( i = 1, \ldots, I \) modified items. Modified items refer to the items in the original assessment that are chosen for modification. In other words, they are the “source” or “parent” items (from the original assessment) that are modified for Braille
Table 3.1: Data Structure for Original Calibration Data

<table>
<thead>
<tr>
<th>Item</th>
<th>Unmodified</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>...</td>
<td>J</td>
</tr>
<tr>
<td>Person</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td></td>
<td>y_{pj}</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Braille Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Test items. Unmodified items are the remaining ones that are not chosen for modification. Let $y_{pj} \in \{0, 1\}$ denote the response from person $p$ to unmodified item $j$ and $y_{pi} \in \{0, 1\}$ the response from person $p$ to modified item $i$.

Table 3.1 presents the structure of the item response data. For convenience, we group unmodified items in one block and modified items in another to form two separate blocks. Each block is essentially a cross-classification of the person-by-item design, i.e., a crossed $p \times j$ design and $p \times i$ design. Specifically in the context of ELPA21, $y_{pj}$ and $y_{pi}$ in the $P \times (J + I)$ matrix represent the responses to the original assessment items from the general EL population. The design displays the case where some items from the original assessment are modified while some items are not (or cannot) be modified. The bottom block contains the Braille test items to be presented to BLV students, and these Braille items are of our interest. As shown to the left of the Braille item block, there is one-to-one correspondence between modified items and the modified Braille items, as indicated by the indices $i = 1, ..., I$. To the right of the block lie newly developed Braille test items ($n = 1, ..., N$) – these obviously have no counterparts in the original assessment. In the following discussion, we only focus on the modified Braille items, i.e., the ones whose source are the modified items.

The next is cross-classification of the judgmental data on the modified Braille items.
### Table 3.2: Data Structure for Judgmental Data

<table>
<thead>
<tr>
<th>Rater</th>
<th>1</th>
<th>...</th>
<th>i</th>
<th>...</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>y_{ri1}</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>y_{rik}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The raters \((r = 1, ..., R)\) provide \(k = 1, ..., K\) ratings for each Braille test item \(i\). Let \(y_{rik} \in [0, 1]\) denote the \(k\)th rating in probability scale from rater \(r\) to item \(i\). It is important to note that the raters are informed of the proportion of correct responses to the source items from the general EL population. Table 3.2 shows the structure of the judgmental data. It is a two-way table with raters on rows and items on columns, i.e., a crossed \(r \times i\) design, having repeated measures in each cell. In ELPA21, for example, raters provided ratings on two cut-scores: the lower cut-score (the Level 3 cut-score, or the scale score separating Level 2 and 3) and the upper cut-score (the Level 4 cut-score, or the scale score separating Level 3 and 4), on \(I\) modified Braille test items. It yields a bivariate structure with two outcomes in each cell \((K = 2)\). Each cell would have ratings at those two cut-scores, \(y_{1ri}\) and \(y_{2ri}\), for Braille test item \(i\). To reiterate, these represent the expected probability of correct response as judged by a rater at the lower and upper cut-scores.

### 3.2 Model Formulation

We are looking eventually to construct a three-part model: two from the original calibration data and one from the judgmental data. Our goal is to revise item parameters for the modified Braille test items in an alternative manner given that direct calibration of these items is not feasible. The judgmental data on the Braille items and its model therefore serve as the backbone for moderated item calibration. Before introducing the full
three-part model, however, for simplicity we start by first supposing that there is no original calibration data set and model the judgmental data without it.

Assuming that the raters only characterize items by difficulty, the variance in raters’ judgment can be split into two components: the variability in item difficulty and the variability in raters. This is a set-up for a CCREM. The model is nonlinear after the non-linear nature of IRT from which the item parameters are drawn, and multiple judgments in each cell render the model multivariate. This amounts to a multivariate nonlinear cross-classified random effects model. To aid understanding, we present the model as a two-level hierarchical model: “Within-cell” model and “Between-cell” model, following Raudenbush and Bryk (2002).

"Within-Cell" Model. The ratings are nested within each cell, and in each cell, we have repeated measures of ratings. Recall that \( y_{rik} \in [0, 1] \) is the kth rating in probability scale from rater r on item i. At cut-score \( k \), denoted \( \theta^*_k \), the unconditional within-cell model can be written as

\[
\begin{bmatrix}
  y_{r1} \\
  \vdots \\
  y_{rK}
\end{bmatrix} =
\begin{bmatrix}
  \frac{1}{1+\exp[-(\theta^*_1 + \eta_{rik})]} \\
  \vdots \\
  \frac{1}{1+\exp[-(\theta^*_K + \eta_{rik})]}
\end{bmatrix} +
\begin{bmatrix}
  e_{r1} \\
  \vdots \\
  e_{rK}
\end{bmatrix}, \quad (3.1)
\]

where \( \eta_{rik} \) is the mean of the kth rating in cell ri, and \( e_{rik} \) is the random error and/or interaction effect between rater and item. Collectively, \( e_{ri} = (e_{r1}, ..., e_{rK})' \sim N_K(0, \Sigma_e) \), where

\[
\Sigma_e =
\begin{pmatrix}
  \sigma^2_{e1} & \sigma_{e12} & \cdots & \sigma_{e1K} \\
  \sigma_{e21} & \sigma^2_{e2} & \cdots & \sigma_{e2K} \\
  \vdots & \vdots & \ddots & \vdots \\
  \sigma_{eK1} & \sigma_{eK2} & \cdots & \sigma^2_{eK}
\end{pmatrix}.
\]

It is important to note that while the model resembles an IRT model, we have a fixed value of cut-score \( \theta^*_k \) because the probability of correct response, \( y_{rik} \), was obtained at a particular \( \theta^* \).
"Between-Cell" Model. Next, we model variations among ratings (between cells) by splitting them into item and rater components. The unconditional between-cell model can be formulated as

\[
\eta_{ri1} = \mu + \alpha_i + \gamma_r \\
\vdots \\
\eta_{riK} = \mu + \alpha_i + \gamma_r,
\]

where \( \mu \) is the fixed intercept (the overall mean) of all ratings, \( \alpha_i \sim N(0, \sigma^2_\alpha) \) is the random item effect, and \( \gamma_r \sim N(0, \sigma^2_\gamma) \) is the random rater effect. Unlike in the illustration of CCREM given in Chapter 2, the sources of variation of \( y_{rik} \) are item and rater (as opposed to item and person) due to the fact it is given from raters’ judgment on items, not from persons’ responses. Still, the model representation is no different from that in Chapter 2.

Combined Model. Equation 3.2 and Equation 3.1 together lead to the combined unconditional multivariate nonlinear CCREM:

\[
\begin{bmatrix}
    y_{r11} \\
    \vdots \\
    y_{riK}
\end{bmatrix} =
\begin{bmatrix}
    \frac{1}{1 + \exp[-(\theta_1^i + \mu + \alpha_i + \gamma_r)]} \\
    \vdots \\
    \frac{1}{1 + \exp[-(\theta_K^i + \mu + \alpha_i + \gamma_r)]}
\end{bmatrix}
\begin{bmatrix}
    e_{r11} \\
    \vdots \\
    e_{riK}
\end{bmatrix}. \tag{3.3}
\]

This multivariate nonlinear CCREM yields revised item parameters for the modified Braille test items using the information from the raters, specifically via the random item effect. In a hypothetical situation where the raters do not have the original calibration data at all and no other source of information is available for modeling, we may have to solely rely on the model above. Then, the new item (difficulty) parameter would be defined as \( \tilde{c}_i = \mu + \alpha_i \). Fortunately in our case, the original calibration data from ELs from the general population is available, and we are able to formulate a three-part model: the Part I and II models for the original calibration data in Table 3.1 and the Part III model for the judgmental data in Table 3.2. We now proceed to formulating the three-part model. Again, without loss of generality, only dichotomous items/ratings are
considered to illustrate the model.

### 3.2.1 Part I

The first model for the unmodified items of the original calibration data, presented by the $p \times j$ design in the left side of Table 3.1, is a two-parameter logistic (2PL) model. If $\theta_p$ be the latent ability for person $p$, then conditional on $\theta_p$, the response of person $p$ to item $j$ is

$$\logit[P(Y_{pj} = 1|\theta_p)] = \logit(\pi_{pj}) = a_j \theta_p + c_j,$$  \hspace{1cm} (3.4)

where $a_j$ and $c_j$ are respectively the item slope parameter and the intercept term for unmodified item $j$. It is assumed that $\theta_p \sim N(0, \sigma^2_\theta)$. For identification, $\sigma^2_\theta$ is usually fixed at 1. We are particularly interested in $\theta_p$ from this model in order to obtain the latent scores of the general EL population. Because the unmodified items do not serve as source items for any Braille test items, the item parameters from this model are not of interest.

### 3.2.2 Part II

The second model is for the modified items of the original calibration data, presented by the $p \times i$ design on the right side of Table 3.1. It is a cross-classified random effects 2PL model. The response of person $p$ to item $i$ is expressed as

$$\logit[P(Y_{pi} = 1|\theta_p, c_i)] = \logit(\pi_{pi}) = a_i^* \theta_p + c_i,$$  \hspace{1cm} (3.5)

where $a_i^*$ is the item slope parameter and $c_i \sim N(\mu_c, \sigma^2_c)$ is the intercept term for modified item $i$. The asterisk punctuates that $a_i^*$ is not estimated but rather fixed to the original calibration slope. This fixation pertains to the aforementioned assumption that raters judge items only by difficulty, and the item parameters of these modified items are directly connected to those of the Braille test items derived from them, which are fully
determined in Part III with ratings. Importantly, $\theta_p$ in Part II is identical to that in Part I because persons are the same between Part I and Part II, that is, the the general ELs.

### 3.2.3 Part III

In this final model, we model raters’ judgment utilizing the information from the original calibration data. We build on the unconditional model (Equations 3.1 to 3.3) by adding item covariate $c_i$, which is estimated in Part II, and obtain a conditional model. There are two principal reasons for adding the item random effect of modified item from Part II model as a covariate. First, the most plausible covariate for predicting the difficulty of the Braille items would be that of their source items. Second, it allows the item parameters derived from raters’ judgments to be comparable to those of the original test items. The model then follows Raudenbush and Bryk (2002)'s formulation.

"Within-Cell" Model. The conditional within-cell model is the same as the unconditional within-cell model (Equation 3.1) except that the offset term $\theta^*_k$ is exchanged for $a_i^*\theta^*_k$ for rescaling purpose. It naturally reflects the 2PL model used in the calibration of the original test items. The within-cell model is then

$$
\begin{bmatrix}
y_{r1} \\
\vdots \\
y_{riK}
\end{bmatrix} =
\begin{bmatrix}
\frac{1}{1+\exp[-(a_i^*\theta^*_1+\eta_{ril})]} \\
\vdots \\
\frac{1}{1+\exp[-(a_i^*\theta^*_K+\eta_{rik})]}
\end{bmatrix}
+ 
\begin{bmatrix}
e_{r1} \\
\vdots \\
e_{riK}
\end{bmatrix}.
$$  

(3.6)

The errors follow the same covariance structure as in the unconditional model.

"Between-Cell" Model. By invoking item covariate ($c_i$ from Equation 3.5), the conditional between-cell model becomes

$$
\eta_{ril} = \mu + \pi c_i + \zeta_i + \gamma_r \\
\vdots \\
\eta_{riK} = \mu + \pi c_i + \zeta_i + \gamma_r,
$$  

(3.7)
where \( \pi \) is the fixed coefficient for \( c_i \) and \( \zeta_i \sim N(0, \sigma^2_\zeta) \) and \( \gamma_i \sim N(0, \sigma^2_\gamma) \) are respectively the residual random item and rater effects after \( c_i \) is taken into account.

Before introducing the full combined model, let us present the “between-cell” model in two separate equations: 1) the unconditional “between-cell” model and 2) the item regression equation or so called the means-as-outcomes regression equation. This way of presentation not only helps understanding of the model and paves a straightforward pathway to the revised new item parameters for the modified Braille test item, but also (and more importantly) has a direct connection to the estimation strategy which will be the topic of Chapter 4. The unconditional between-cell model can then be written as

\[
\begin{align*}
\eta_{ri1} &= \alpha_i + \gamma_r \\
&\vdots \\
\eta_{riK} &= \alpha_i + \gamma_r,
\end{align*}
\]

and the item regression equation as

\[
\alpha_i = \mu + \pi c_i + \zeta_i. \tag{3.9}
\]

It is clear from Equation 3.9 that the revised new item difficulty parameter from the three-part model can be derived as \( \tilde{\alpha}_i = \mu + \pi c_i + \zeta_i \).

**Combined Model.** We can obtain the full combined conditional multivariate nonlinear cross-classified random effects model by plugging in Equation 3.7 into Equation 3.6.

\[
\begin{bmatrix}
y_{ri1} \\
\vdots \\
y_{riK}
\end{bmatrix} = \begin{bmatrix}
\frac{1}{1 + \exp[-(a^* \theta^*_i + \mu + \pi c_i + \zeta_i + \gamma_r)]} \\
\vdots \\
\frac{1}{1 + \exp[-(a^* \theta^*_K + \mu + \pi c_i + \zeta_i + \gamma_r)]}
\end{bmatrix} + \begin{bmatrix}
e_{ri1} \\
\vdots \\
e_{riK}
\end{bmatrix}. \tag{3.10}
\]

At last, we have the full three-part model: 1) a 2PL model or, equivalently, a generalized linear mixed model with fixed item and random person; 2) a cross-classified random
effects 2PL model or, equivalently, a generalized linear mixed model with \textit{random person} and \textit{random item}; 3) a (conditional) multivariate nonlinear \textit{cross-classified random effects} model.
3.3 Likelihood Functions

3.3.1 Part I

For a dichotomous response, the conditional density of $y_{pj}$ is given by the Bernoulli distribution:

$$f(y_{pj} | \theta_p; a_j, c_j) = P(y_{pj} = 1) y_{pj} [1 - P(y_{pj} = 1)]^{1 - y_{pj}}.$$  \hspace{1cm} (3.11)

Let $\mathbf{y}_{pu} = (y_{p1}, ..., y_{pj})'$ be a $J \times 1$ vector of $p$th person’s response pattern on unmodified items. Let $\mathbf{a}_u$ and $\mathbf{c}_u$ be $J \times 1$ vectors of the slope and intercept parameters for all $J$ unmodified items, respectively. Assuming independence conditioning on the latent variable (Lord & Novick, 1968), the conditional density $\mathbf{y}_{pu}$ is

$$f(\mathbf{y}_{pu} | \theta_p; \mathbf{a}_u, \mathbf{c}_u) = \prod_{j=1}^{J} f(y_{pj} | \theta_p; a_j, c_j).$$  \hspace{1cm} (3.12)

With the assumption we specified about the person distribution, the marginal density of $\mathbf{y}_{pu}$ is

$$f(\mathbf{y}_{pu} | \mathbf{a}_u, \mathbf{c}_u) = \int \prod_{j=1}^{J} f(y_{pj} | \theta_p; a_j, c_j) h(\theta_p) d\theta_p,$$  \hspace{1cm} (3.13)

where $h(\theta_p)$ is the standard normal density function of $\theta_p$. This is the marginal density we would obtain with the standard IRT approach.

Let $\mathbf{y}_u$ be a $(P \times J) \times 1$ vector of responses for all $J$ unmodified items from all $P$ persons. The observed data likelihood function is then expressed as

$$L(\omega_1 | \mathbf{y}_u) = \int \prod_{p=1}^{P} \left[ \prod_{j=1}^{J} f(y_{pj} | \theta_p; a_j, c_j) \right] h(\theta_p) d\theta_p,$$  \hspace{1cm} (3.14)

where $\omega_1 = (\mathbf{a}_u', \mathbf{c}_u')'$ is the vector containing all freely estimated parameters in Part I.

The latent variable $\theta_p$ can be treated as missing data. For a $P \times 1$ vector $\theta$ of all
individual latent scores, the complete data likelihood can be defined as

\[ L(\omega_1|y_u, \theta) = \prod_{p=1}^{P} \left[ \prod_{j=1}^{I} f(y_{pj}|\theta_p, a_j, c_j) \right] \left[ \prod_{p=1}^{P} h(\theta_p) \right]. \]  

(3.15)

### 3.3.2 Part II

The conditional density of \( y_{pi} \) is again that of a Bernoulli variable

\[ f(y_{pi}|\theta_p, c_i; a^*_i) = P(y_{pi} = 1)^{y_{pi}}[1 - P(y_{pi} = 1)]^{1-y_{pi}}. \]  

(3.16)

Let \( y_m = (y_{11}, ..., y_{PI})' \) be a \((P \times I) \times 1\) vector of the total response patterns on all \( I \) modified items from all \( P \) persons. Let \( a^*_m \) and \( c_m \) respectively be \( I \times 1 \) vectors of all slope parameters fixed to the original calibration data and intercept parameters for all \( I \) items. Recall that \( \theta \) is the \( P \times 1 \) vector of all persons’ latent scores. Assuming the local person and item independence, the conditional density \( y_m \) is written as

\[ f(y_m|\theta, c_m; a^*_m) = \prod_{i=1}^{I} \prod_{p=1}^{P} f(y_{pi}|\theta_p, c_i; a^*_i). \]  

(3.17)

Let \( \omega_2 = (\mu_c, \sigma^2_c)' \) be the vector of all freely estimated parameters from Part II. Once the distributions of a person and an item are specified, the observed data likelihood function is

\[ L(\omega_2|y_m) = \int \int \prod_{i=1}^{I} \prod_{p=1}^{P} f(y_{pi}|\theta_p, c_i; a^*_i) \left[ \prod_{p=1}^{P} h(\theta_p) \right] \left[ \prod_{i=1}^{I} h(c_i; \mu_c, \sigma^2_c) \right] dc_id\theta_p, \]  

(3.18)

where \( h(c_i; \mu_c, \sigma^2_c) \) is the normal density function of \( c_i \) with mean \( \mu_c \) and variance \( \sigma^2_c \).

Treating the random effect \( \theta_p \) and \( c_i \) as missing data, the complete data likelihood
can be defined as

\[
L(\omega_2 | y_m, \theta, c_m) = \left[ \prod_{i=1}^{l} \prod_{p=1}^{P} f(y_{pi} | \theta_p, c_i; a_i^*) \right] \left[ \prod_{p=1}^{P} h(\theta_p) \right] \left[ \prod_{i=1}^{l} h(c_i; \mu_c, \sigma_c^2) \right]. \tag{3.19}
\]

3.3.3 Part III

Before defining the likelihood for the multivariate nonlinear CCREM of the judgmental data, let us write the model in a matrix form of NLME model introduced in Section 2.3.3. Let \( y_{ri} = (y_{ri1}, ..., y_{riK})' \) denote a \( K \times 1 \) vector of ratings for \( K \) ratings in a cell. In addition, let \( e_{ri} = (e_{ri1}, ..., e_{riK})' \) be a corresponding \( K \times 1 \) vector of the error terms. First, we define an unconditional model by substituting Equation 3.8 into Equation 3.6. It can be expressed in the matrix form as

\[
y_{ri} = g(\phi_{ri}) + e_{ri}, \tag{3.20}
\]

where \( g(\cdot) \) indicates a vector of nonlinear differentiable function of mixed effects parameter vector \( \phi_{ri} \), which is defined as

\[
g(\phi_{ri}) = \begin{bmatrix}
\frac{1}{1+\exp[-(a_{i1}^i\theta_{1i}+a_i^r+\gamma_r)]} \\
\vdots \\
\frac{1}{1+\exp[-(a_i^K\theta_{Ki}+a_i^r+\gamma_r)]}
\end{bmatrix}. \tag{3.21}
\]

Now let \( \omega_3 = (\mu, \pi, \sigma_{\xi}^2, \sigma_{\gamma}^2, \text{vech}(\Sigma_e))' \) be a vector collecting all freely estimated parameters. We split \( \omega_3 \) into \( \omega_{3,1} = (\sigma_{\xi}^2, \text{vech}(\Sigma_e))' \) and \( \omega_{3,2} = (\mu, \pi, \sigma_{\gamma}^2) \).

The conditional density of \( y_{ri} \) is the product of the density of \( y_{ri} \) from the uncondi-
tional model and the density of $\alpha_i$ from item regression (Equation 3.9):

$$f(y_{ri} | \alpha_i, \gamma_r; \omega_3) = f(y_{ri} | \alpha_i, \gamma_r; \omega_{3,1}) f(\alpha_i | \omega_{3,2})$$

$$= \left| 2\pi \Sigma_e \right|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y_{ri} - g(\phi_{ri}))' \Sigma_e^{-1} (y_{ri} - g(\phi_{ri})) \right\}$$

$$\times \frac{1}{\sqrt{2\pi \sigma_\zeta^2}} \exp \left\{ -\frac{(\alpha_i - x' \beta)^2}{2\sigma_\zeta^2} \right\},$$

(3.22)

where $x = (1, c_i)'$ and $\beta = (\mu, \pi)'$. Let $\hat{y} = (y_{111}, ..., y_{RIK})'$ be an $(R \times I \times K) \times 1$ vector of total ratings on all $I$ modified (Braille) test items by all $R$ raters on $K$ cut-scores. Let $u_{\alpha} = (\alpha_1, ..., \alpha_I)'$ be an $I \times 1$ vector of random item effects and $u_{\gamma} = (\gamma_1, ..., \gamma_I)'$ be an $R \times 1$ vector of random rater effects. The conditional density of $\hat{y}$ is then

$$f(\hat{y} | u_\gamma; \omega_3) = f(\hat{y} | u_{\alpha}, u_\gamma; \omega_{3,1}) f(u_{\alpha} | \omega_{3,2})$$

$$= \prod_{i=1}^{I} \prod_{r=1}^{R} f(y_{ri} | \alpha_i, \gamma_r; \omega_{3,1}) \times \prod_{i=1}^{I} f(\alpha_i | \omega_{3,2}).$$

(3.23)

Once $\gamma_r$ and $\alpha_i$ are integrated out, the observed data likelihood of $\hat{y}$ is obtained as follows:

$$L(\omega_3 | \hat{y}) = \int \int \prod_{i=1}^{I} \prod_{r=1}^{R} f(y_{ri} | \alpha_i, \gamma_r; \omega_{3,1}) \left[ \prod_{r=1}^{R} h(\gamma_r; \sigma_\gamma^2) \right] \left[ \prod_{i=1}^{I} f(\alpha_i | \omega_{3,2}) \right] d\alpha_i d\gamma_r$$

(3.24)

where $h(\gamma_r; \sigma_\gamma^2)$ is the normal density function of $\gamma_r$. Here, we treat the random effects $\gamma_r$ and $\alpha_i$ as missing data. By augmenting $\gamma_r$ and $\alpha_i$, the complete data likelihood is written as

$$L(\omega_3 | \hat{y}, u_{\alpha}, u_\gamma) = \left[ \prod_{i=1}^{I} \prod_{r=1}^{R} f(y_{ri} | \alpha_i, \gamma_r; \omega_{3,1}) \right] \left[ \prod_{r=1}^{R} h(\gamma_r; \sigma_\gamma^2) \right] \left[ \prod_{i=1}^{I} f(\alpha_i | \omega_{3,2}) \right].$$

(3.25)
3.3.4 The Joint Likelihood

Let $Y$ be the set of observed data $\{y_u, y_m, \tilde{y}\}$ and $M$ be the set of missing data $\{\theta, c_m, u_\alpha, u_\gamma\}$. The joint likelihood function for the three-part model is the product of the three likelihood functions:

$$L(\omega|Y,M) = L(\omega_1|y_u, \theta)L(y_m|c_m)L(\omega_3|\tilde{y}, u_\alpha, u_\gamma), \quad (3.26)$$

where $\omega = (\omega_1', \omega_2', \omega_3')'$. Taking the logarithm, the joint log-likelihood function for three-part model becomes the sum of the three log-likelihood functions:

$$l(\omega|Y,M) = l(\omega_1|y_u, \theta) + l(\omega_2|y_m, c_m) + l(\omega_3|\tilde{y}, u_\alpha, u_\gamma). \quad (3.27)$$
CHAPTER 4

Estimation and Scoring for Cross-classified Random Effects Model

This chapter is on the application of the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b) to the estimation of the unified CCREM for moderated item calibration. MH-RM is aimed at efficient and accurate recovery of the maximum likelihood (ML) solution of the model parameters. First, MH-RM is introduced and defined, followed by outlines of the MH sampler implementation and the necessary ingredients for the RM update within the context of the proposed model. Next, we show how standard errors (SEs) are obtained under MH-RH. Then, scoring of random effects using the empirical Bayes (EB) prediction is described, which finally leads to the revised item parameter estimates. Miscellaneous estimation strategies that may be helpful for implementing the algorithm for the proposed model are provided as well.

4.1 Model Estimation via the MH-RM Algorithm

MH-RM is a data-augmented RM algorithm (Robbins & Monro, 1951) coupled with the MH algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). This fusion was proposed by Cai (2008, 2010a, 2010b) based on two critical insights, which are so fundamental to the definition and implementation of the algorithm that it is worth directly quoting Cai and Thissen (2015):

*The Monte Carlo average of complete data log-likelihood gradients gives the same likelihood ascent direction as the observed data log-likelihood gradient vector (p.50).*
In Robbins and Monro’s context, they were attempting to find roots of noise-corrupted regression functions, where the noise may be due to observational measurement error; in our context, we purposefully inject Monte Carlo noise by imputing the missing data (the latent variable scores), so that we can observe an approximate direction of likelihood ascent (p.50).

The first insight is rooted in Fisher’s (1925) identity, and the second is rooted in Robbins and Monro’s (1951) classical Stochastic Approximation (SA) method.

Let $\omega^{(t)}$ be the parameter estimates at iteration $t$. At iteration $t + 1$, the MH-RM algorithm follows three steps: stochastic imputation, stochastic approximation, and the RM update.

**Step 1: Stochastic Imputation**

Draw $m_t$ sets of missing data $\{M_j^{(t+1)}; j = 1, ..., m_t\}$ using the MH sampler with the posterior predictive distribution $\Pi(M|Y, \omega^{(t)})$ of missing data. Then $m_t$ sets of complete data $\{Y, M_j^{(t+1)}; j = 1, ..., m_t\}$ are created. Details of implementing the MH sampler can be found in next section (Section 4.1.2).

**Step 2: Stochastic Approximation**

Based on the imputed data, the ascent directions towards $\omega^{(t+1)}$ are determined by evaluating the complete data log-likelihood and its gradients. The gradient of the complete data log-likelihood is

$$s(\omega^{(t)}|Y, M_j^{(t+1)}) = \nabla_{\omega} l(\omega^{(t)}|Y, M_j^{(t+1)}), \quad (4.1)$$

where the $\nabla_{\omega}$ operator returns a vector of first order derivatives with respect to $\omega$. In practice, we can approximate it with the sample average of the complete data gradients:

$$\tilde{s}_{t+1} = \frac{1}{m_t} \sum_{j=1}^{m_t} s(\omega^{(t)}|Y, M_j^{(t+1)}). \quad (4.2)$$

This procedure is made possible by Fisher’s identity (Fisher, 1925), which states that
the conditional expectation of the gradient of the complete data log-likelihood over the posterior distribution of missing data is equal to the gradient of the observed data log-likelihood:

$$\nabla_\omega l(\omega|Y) = \int s(\omega|Y,M)\Pi(M|Y,\omega)dM.$$  \hspace{1cm} (4.3)

In addition, we compute the conditional expectation of the complete data information matrix, the purpose of which is to improve stability and speed. For complete data information matrix

$$H(\omega|Y,M) = -\frac{\partial^2 l(\omega|Y,M)}{\partial \omega \partial \omega'},$$  \hspace{1cm} (4.4)

we can also compute its Monte Carlo approximation as

$$H_{t+1} = \frac{1}{m_t} \sum_{j=1}^{m_t} H(\omega^{(t)}|Y,M^{(t+1)}_j).$$  \hspace{1cm} (4.5)

**Step 3: The Robbins-Monro Update**

The RM filter is applied as we update the parameter estimates. It is

$$\omega^{(t+1)} = \omega^{(t)} + \epsilon_t (\Gamma^{-1}_{t+1}) \bar{s}_{t+1}$$  \hspace{1cm} (4.6)

where $\Gamma^{-1}_{t+1}$ is a recursive stochastic approximation of the conditional expectation of the complete data information matrix. Given initial values $(\omega_0, \Gamma_0)$, where $\Gamma_0$ is a symmetric positive definite matrix, a recursive approximation of $E(H(\omega|Y,M))$ is defined as

$$\Gamma_{t+1} = \Gamma_t + \epsilon_t (H_{t+1} - \Gamma_t),$$  \hspace{1cm} (4.7)

where $\epsilon_t$ is a sequence of gain constants such that $\epsilon_t \in (0,1]$, subject to $\sum_{t=0}^{\infty} \epsilon_t = \infty$ and $\sum_{t=0}^{\infty} \epsilon_t^2 < \infty$. Essentially, the role of $\epsilon_t$ is to eliminate the noise effect introduced by imputing for missing data.

Often in practice, the iterations are divided into multiple stages (Cai, 2015; Falk & Cai, 2016; Monroe & Cai, 2014; Monroe, 2014; Yang & Cai, 2014). In Stage I, we run
some iterations to bring the starting values to the general neighborhood of the parameter estimates. Further iterations are subsequently performed in Stage II, wherein the averages of the estimates in the neighborhood become the starting values for in Stage III. Accordingly, multi-stage gain constants are favored for $\epsilon_t$ such that gain constants with fixed values are set in Stage I and Stage II, followed by decreasing gain constants in Stage III (Cai, 2008, 2015). Note that the gain constants can be determined after monitoring the traces of parameters. Iterations over the three steps is terminated when the minimum successive differences of the estimates for a predetermined window size reaches a convergence criteria.

4.1.1 Implementation of the MH Sampler

Recall that we need to draw plausible values of $M_j(t+1)$ from its posterior predictive distribution of missing data $\Pi(M|y, \omega(t))$ in Step 1 of the MH-RM workflow. To impute $M$, a Metropolis-within-Gibbs algorithm by Patz and Junker (1999) shall be used, following Cai (2008, 2010a, 2010b).

First, let $f(\theta_1|\theta_1, ..., \theta_{p-1}, \theta_{p+1}, ..., \theta_p, Y, c_m, u, u, \omega)$ be the full conditional density for $\theta_p$. Let $\theta^t_p$ denote the value of $\theta_p$ in the $t$-th iteration of Gibbs sampler. The Gibbs sampling algorithm can be constructed as follows:

\[
\begin{align*}
\text{Draw } \theta^t_1 & \sim f(\theta^t_1|\theta^{(t-1)}_1, ..., \theta^{(t-1)}_p, Y, c_m, u, u, \omega) \\
\text{Draw } \theta^t_2 & \sim f(\theta^t_2|\theta^t_1, \theta^{(t-1)}_3, ..., \theta^{(t-1)}_p, Y, c_m, u, u, \omega) \\
& \vdots \\
\text{Draw } \theta^t_p & \sim f(\theta^t_p|\theta^t_1, ..., \theta^t_{p-1}, \theta^{(t-1)}_{p+1}, ..., \theta^{(t-1)}_p, Y, c_m, u, u, \omega) \\
& \vdots \\
\text{Draw } \theta^t_P & \sim f(\theta^t_P|\theta^t_1, ..., \theta^t_{P-1}, Y, c_m, u, u, \omega).
\end{align*}
\]

Next, let $f(c_i|c_1, ..., c_{i-1}, c_{i+1}, ..., c_I, Y, \theta, u, u, \omega)$ be the full conditional density for $c_i$. 

44
Let $c^t_i$ denote the value of $c_i$ in the $t$-th iteration of a Gibbs sampler. In the same manner, the Gibbs sampling algorithm can be constructed as follows:

\[
\begin{align*}
\text{Draw } c^t_1 & \sim f(c^t_1 | c^t_2, \ldots, c^t_i, Y, \theta, u_\gamma, \omega) \\
\text{Draw } c^t_2 & \sim f(c^t_2 | c^t_1, c^t_3, \ldots, c^t_i, Y, \theta, u_\gamma, \omega) \\
& \quad \vdots \\
\text{Draw } c^t_i & \sim f(c^t_i | c^t_1, \ldots, c^t_{i-1}, c^t_{i+1}, \ldots, c^t_i, Y, \theta, u_\gamma, \omega) \\
& \quad \vdots \\
\text{Draw } c^t_I & \sim f(c^t_I | c^t_1, \ldots, c^t_{I-1}, Y, \theta, u_\gamma, \omega).
\end{align*}
\] (4.9)

Similarly, let $f(\alpha_i | \alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_I, Y, \theta, c_m, u_\gamma, \omega)$ be the full conditional density for $\alpha_i$, and let the value of $\alpha_i$ in the $t$-th iteration of a Gibbs sampler be denoted $\alpha^t_i$. The Gibbs sampling algorithm consists of the following steps:

\[
\begin{align*}
\text{Draw } \alpha^t_1 & \sim f(\alpha^t_1 | \alpha^t_2, \ldots, \alpha^t_I, Y, \theta, c_m, u_\gamma, \omega) \\
\text{Draw } \alpha^t_2 & \sim f(\alpha^t_2 | \alpha^t_1, \alpha^t_3, \ldots, \alpha^t_I, Y, \theta, c_m, u_\gamma, \omega) \\
& \quad \vdots \\
\text{Draw } \alpha^t_i & \sim f(\alpha^t_i | \alpha^t_1, \ldots, \alpha^t_{i-1}, \alpha^t_{i+1}, \ldots, \alpha^t_I, Y, \theta, c_m, u_\gamma, \omega) \\
& \quad \vdots \\
\text{Draw } \alpha^t_I & \sim f(\alpha^t_I | \alpha^t_1, \ldots, \alpha^t_{I-1}, Y, \theta, c_m, u_\gamma, \omega).
\end{align*}
\] (4.10)

Lastly, let $f(\gamma_r | \gamma_1, \ldots, \gamma_{r-1}, \gamma_{r+1}, \ldots, \gamma_R, Y, \theta, c_m, u_\gamma, \omega)$ be the full conditional density for $\gamma_r$, and let $\gamma^t_r$ be the value of $\gamma_r$ in the $t$-th iteration of a Gibbs sampler. Again, the
Gibbs sampling algorithm follows

\[
\begin{align*}
\text{Draw} & \quad \gamma_t^1 \sim f(\gamma_1 | \gamma_2^{(t-1)}, \ldots, \gamma_R^{(t-1)}, Y, \theta, c_m, u, \omega) \\
\text{Draw} & \quad \gamma_t^2 \sim f(\gamma_2 | \gamma_1^t, \gamma_3^{(t-1)}, \ldots, \gamma_R^{(t-1)}, Y, \theta, c_m, u, \omega) \\
& \quad \vdots \\
\text{Draw} & \quad \gamma_t^r \sim f(\gamma_r | \gamma_1^t, \ldots, \gamma_{r-1}^t, \gamma_{r+1}^t, \ldots, \gamma_R^{(t-1)}, Y, \theta, c_m, u, \omega) \\
& \quad \vdots \\
\text{Draw} & \quad \gamma_t^R \sim f(\gamma_R | \gamma_1^t, \ldots, \gamma_{R-1}^t, Y, \theta, c_m, u, \omega).
\end{align*}
\] (4.11)

Even with these full conditional densities for \(\theta, c, \alpha,\) and \(\gamma\) values, direct sampling from the distributions is still not feasible. That being said, the Gibbs sampling must be coupled with the MH algorithm.

First, to draw each \(\theta_p\), we start with an initial value \(\theta_p^0\). At iteration \(t + 1\), given the current state \(\theta_p^t\), we simulate a candidate \(\theta_p^k\) from a transition kernel (or proposal distribution) \(q(\theta_p^k | \theta_p^t)\). After canceling out the densities that are not dependent on \(\theta_p\) from the joint densities of the three-part model, we can compute the likelihood ratio as

\[
R(\theta_p^k, \theta_p^t) = \frac{f(y_{pu} | \theta_p^k; \omega_1) f(y_{pm} | \theta_p^k; \omega_2) h(\theta_p^k) q(\theta_p^k)}{f(y_{pu} | \theta_p^t; \omega_1) f(y_{pm} | \theta_p^t; \omega_2) h(\theta_p^t) q(\theta_p^t)}.
\] (4.12)

The symmetry of the proposal distribution, that is \(q(\theta_p^k, \theta_p^t) = q(\theta_p^t, \theta_p^k)\), allows us to simplify the ratio to

\[
R(\theta_p^k, \theta_p^t) = \frac{f(y_{pu} | \theta_p^k; \omega_1) f(y_{pm} | \theta_p^k; \omega_2) h(\theta_p^k)}{f(y_{pu} | \theta_p^t; \omega_1) f(y_{pm} | \theta_p^t; \omega_2) h(\theta_p^t)}.
\] (4.13)

The acceptance probability of moving from state \(\theta_p^t\) to \(\theta_p^k\) is \(P = \min(R, 1)\). When \(P = 1\), we accept the candidate value and set \(\theta_p^{t+1} = \theta_p^k\). When \(P = R < 1\), we reject the candidate value and remain in the current state \(\theta_p^{t+1} = \theta_p^t\).

We carry the same logic over to the drawing of \(c_i, \alpha_i\) and \(\gamma_r\). After simplifications,
the likelihood ratio for $c_i$ is

$$R(c^k_i,c^l_i) = \frac{f(y_i|c^k_i; \omega_2)h(c^k_i|\mu_c,\sigma^2_c)f(\alpha_i|\omega_{3,2})}{f(y_i|c^l_i; \omega_2)h(c^l_i|\mu_c,\sigma^2_c)f(\alpha_i|\omega_{3,2})}. \quad (4.14)$$

Note that $f(\alpha_i|\omega_{3,2})$ is also dependent on the covariate $c_i$. And the following are the corresponding ratios for $\alpha_i$ and $\gamma_r$ after simplifications:

$$R(\alpha^k_i,\alpha^l_i) = \frac{f(\tilde{y}_i|\alpha^k_i,\gamma_r; \omega_{3,1})f(\alpha^k_i|\omega_{3,2})}{f(\tilde{y}_i|\alpha^l_i,\gamma_r; \omega_{3,1})f(\alpha^l_i|\omega_{3,2})}, \quad (4.15)$$

$$R(\gamma^k_r,\gamma^l_r) = \frac{f(\tilde{y}_r|\alpha_i,\gamma^k_r; \omega_{3,1})h(\gamma^k_r|\sigma^2_{\gamma})}{f(\tilde{y}_r|\alpha_i,\gamma^l_r; \omega_{3,1})h(\gamma^l_r|\sigma^2_{\gamma})}. \quad (4.16)$$

For each of the proposal draw, a random walk chain can be used. In other words, a random perturbation (error) is added to the current state as the proposal. Specifically, $\theta^*_i = \theta_i + \delta_i$, $c^*_i = c_i + \delta_i$, $\alpha^*_i = \alpha_i + \delta_i$, and $\gamma^*_i = \gamma_i + \delta_i$, where the incremental density $\delta_i \sim N(0,d^2)$, with the dispersion parameter $d$. A variant of choosing $\delta_i$ would be to set its standard deviation as $d$ multiplied by the standard deviation of the missing data, which is updated every iteration. The choice of $d$ is determined with the goal of maintaining the acceptance ratio of the MH chain. The random-walk Metropolis is commonly used in practice because in many cases it is not easy to come up with a good proposal distribution.

It is important to point out that sampling of $\theta$ values can be done in parallel as the persons are considered independent. The same is true of $c$, $\alpha$, and $\gamma$ values by analogy. In turn, sampling $\theta$, $c_m$, $u_\alpha$, and $u_\gamma$ are done in alternation. One may have noticed that CCREM can be easily implemented with MH-RM in principle, putting aside subtleties. A key idea is imputing each side of missing data (random effects) treating the other side as known by filling in. In a broad sense, the strategy is similar to the imputation posterior (IP) or alternating imputation posterior (AIP) algorithm (Cho & Rabe-Hesketh, 2011; Clayton & Rasbash, 2011).
4.1.2 Complete Data Models and Derivatives

4.1.2.1 Part I

Suppressing subscripts, we rewrite the response of a person to an item in Equation 3.4 in the form prior to the use of logit link:

\[ P = \frac{1}{1 + \exp[-(a\theta + c)]]}. \quad (4.17) \]

The individual case contribution to the complete data log-likelihood can be obtained as

\[ l = y\log P + (1 - y)\log(1 - P). \quad (4.18) \]

The first derivatives of Equation 4.18 are

\[ \frac{\partial l}{\partial a} = \frac{\partial l}{\partial P} \frac{\partial P}{\partial a} = \left( \frac{y}{P} - \frac{1 - y}{1 - P} \right) (1 - P)P\theta = (y - P)\theta \]

\[ \frac{\partial l}{\partial c} = \frac{\partial l}{\partial P} \frac{\partial P}{\partial c} = \left( \frac{y}{P} - \frac{1 - y}{1 - P} \right) (1 - P)P = y - P \]

The second derivatives of Equation 4.18 are

\[ \frac{\partial^2 l}{\partial c^2} = \left( \frac{\partial}{\partial c} \frac{\partial l}{\partial P} \right) \frac{\partial P}{\partial c} + \frac{\partial l}{\partial P} \left( \frac{\partial}{\partial c} \frac{\partial P}{\partial c} \right), \]

\[ \frac{\partial^2 l}{\partial a^2} = \left( \frac{\partial}{\partial a} \frac{\partial l}{\partial P} \right) \frac{\partial P}{\partial a} + \frac{\partial l}{\partial P} \left( \frac{\partial}{\partial a} \frac{\partial P}{\partial a} \right), \]

\[ \frac{\partial^2 l}{\partial a \partial c} = \left( \frac{\partial}{\partial a} \frac{\partial l}{\partial P} \right) \frac{\partial P}{\partial c} + \frac{\partial l}{\partial P} \left( \frac{\partial}{\partial a} \frac{\partial P}{\partial c} \right). \]
where

\[ \frac{\partial}{\partial c} \frac{\partial l}{\partial P} = \left( -\frac{y}{P^2} - \frac{(1-y)}{(1-P)^2} \right) \frac{\partial P}{\partial c}, \]

\[ \frac{\partial}{\partial a} \frac{\partial l}{\partial P} = \left( -\frac{y}{P^2} - \frac{(1-y)}{(1-P)^2} \right) \frac{\partial P}{\partial a}, \]

\[ \frac{\partial}{\partial c} \frac{\partial P}{\partial c} = P(1-P)(1-2P), \]

\[ \frac{\partial}{\partial a} \frac{\partial P}{\partial a} = P(1-P)(1-2P)\theta^2, \]

\[ \frac{\partial}{\partial a} \frac{\partial P}{\partial c} = P(1-P)(1-2P)\theta. \]

### 4.1.2.2 Part II

The complete data model we need from Part II is only the normal model. The contribution of an item to the complete data log-likelihood is derived as

\[ l = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_c^2) - \frac{1}{2} \left( \frac{c_i - \mu_c}{\sigma_c^2} \right)^2. \]  

(4.19)

The first derivatives of Equation 4.19 are

\[ \frac{\partial l}{\partial \mu} = \frac{c_i - \mu}{\sigma_c^2}, \]

\[ \frac{\partial l}{\partial \sigma_c^2} = -\frac{1}{2\sigma_c^2} + \frac{(c_i - \mu)^2}{2\sigma_c^4}. \]  

(4.20)

The second derivatives are

\[ \frac{\partial^2 l}{\partial \mu \partial \mu} = -\frac{1}{\sigma_c^2}, \]

\[ \frac{\partial^2 l}{\partial \mu \partial \sigma_c^2} = -\frac{c_i - \mu}{\sigma_c^4}, \]

\[ \frac{\partial^2 l}{\partial \sigma_c^2 \partial \sigma_c^2} = -\frac{(c_i - \mu)^2}{2\sigma_c^6}. \]  

(4.21)
4.1.2.3 Part III

Let the collection of the function of nonlinear mixed-effects parameter vector be \( g(\phi) = (g(\phi_{11})', g(\phi_{12})', ..., g(\phi_{RI})')' \). Let \( \Phi = I_{R \times I} \otimes \Sigma_e \) be the \( N \times N \) error covariance matrix. Let \( X = [I_t|c_m] \) be the design matrix of fixed effects. The fixed effects are collected in \( \beta = (\mu, \pi)' \). The complete data log-likelihood can be written as

\[
\begin{align*}
l(\omega_3|\tilde{y}, u_a, u_\gamma) &= l(\omega_{3,1}|\tilde{y}, u_a, u_\gamma) + l(\omega_{3,2}|u_a) \\
\alpha &= -\frac{1}{2} \log |\Phi| - \frac{1}{2} \left[ (\tilde{y} - g(\phi))' \Phi^{-1}(\tilde{y} - g(\phi)) \right] - R \frac{1}{2} \log(\sigma_\gamma^2) - \frac{1}{2} \frac{u_\gamma'(u_\gamma - X\beta)}{\sigma_\gamma^2} \\
&= -\frac{1}{2} \log(\sigma_\xi^2) - \frac{1}{2} \frac{u_a'(u_a - X\beta)}{\sigma_\xi^2}.
\end{align*}
\]

(4.22)

The first derivatives of Equation 4.22 with respect to \( \beta = (\mu, \pi)' \), \( \sigma_\xi^2 \), and \( \sigma_\gamma^2 \) are

\[
\begin{align*}
\frac{\partial l(\omega_3|\tilde{y}, u_a, u_\gamma)}{\partial \beta} &= X' u_a - X' X \beta \\
\frac{\partial l(\omega_3|\tilde{y}, u_a, u_\gamma)}{\partial \sigma_\xi^2} &= -I \frac{1}{2} - \frac{1}{2} \frac{(u_a - X\beta)'(u_a - X\beta)}{\sigma_\xi^4} \\
\frac{\partial l(\omega_3|\tilde{y}, u_a, u_\gamma)}{\partial \sigma_\gamma^2} &= -R \frac{1}{2} - \frac{1}{2} \frac{u_\gamma'(u_\gamma - X\beta)}{\sigma_\gamma^4}.
\end{align*}
\]

(4.23)

The second derivative of Equation 4.22 of the complete data log-likelihood are

\[
\begin{align*}
\frac{\partial^2 l(\omega_3|\tilde{y}, u_a, u_\gamma)}{\partial \beta \partial \beta'} &= -X'X \\
\frac{\partial^2 l(\omega_3|\tilde{y}, u_a, u_\gamma)}{\partial \sigma_\xi^2 \partial \sigma_\xi^2} &= I \frac{1}{2} - \frac{1}{2} \frac{(u_a - X\beta)'(u_a - X\beta)}{\sigma_\xi^4} \\
\frac{\partial^2 l(\omega_3|\tilde{y}, u_a, u_\gamma)}{\partial \sigma_\gamma^2 \partial \sigma_\gamma^2} &= R \frac{1}{2} - \frac{1}{2} \frac{u_\gamma'(u_\gamma - X\beta)}{\sigma_\gamma^4}.
\end{align*}
\]

(4.24)

In addition, let \( \phi_i \) and \( \phi_j \) be the \( i \)th and \( j \)th elements of the parameters in \( \Phi \), respectively.
The first derivative of Equation 4.22 with respect to $\phi_i$ is

$$
\frac{\partial l(\omega_3|\tilde{y}, u, u')}{\partial \phi_i} = -\frac{1}{2} \left\{ \text{tr} \left( \Phi_i^{-1} \frac{\partial \Phi}{\partial \phi_i} \right) - (y - g(\phi))' \Phi_i^{-1} \frac{\partial \Phi}{\partial \phi_i} \Phi_i^{-1} (y - g(\phi)) \right\},
$$

(4.25)

and the normal second derivatives with respect to $\phi_i$ and $\phi_j$ are

$$
\frac{\partial^2 l(\omega|\tilde{y}, u, u')}{\partial \phi_j \partial \phi_i} = -\frac{1}{2} \left\{ \text{tr} \left( -\Phi_j^{-1} \frac{\partial \Phi}{\partial \phi_j} \Phi_i^{-1} \frac{\partial \Phi}{\partial \phi_i} + \Phi_j^{-1} \frac{\partial^2 \Phi}{\partial \phi_j \partial \phi_i} \right) 
\right. 
\left. - (y - g(\phi))' \left[ (-1) \Phi_j^{-1} \frac{\partial \Phi}{\partial \phi_j} \Phi_i^{-1} \frac{\partial \Phi}{\partial \phi_i} \Phi_j^{-1} + \Phi_j^{-1} \frac{\partial^2 \Phi}{\partial \phi_j \partial \phi_i} \Phi_j^{-1} \right] (y - g(\phi)) \right\}.
$$

(4.26)

Any crossed second derivative is 0.

### 4.2 Estimation of Standard Errors

Standard errors (SEs) are also available under MH-RM. Two types of SEs, named after the methods used to obtain them, have been proposed: 1) recursively approximated standard errors and 2) post-convergence approximated standard errors (Yang & Cai, 2014). SEs are generally obtained by the square-root of the diagonal elements of the inverse of the observed data information matrix. Both approaches follow Louis (1982) for approximating the observed data information matrix. They have been used and examined in a number of research studies for various models (e.g., Cai, 2010a; Yang & Cai, 2014; Falk & Cai, 2016; Monroe & Cai, 2014). According to Louis (1982), the information matrix of the observed data log-likelihood is

$$
-\frac{\partial^2 l(\omega|Y)}{\partial \omega \partial \omega'} = E(H(\omega|Y, M)) - E(s(\omega|Y, M)|s(\omega|Y, M)')' 
+ E(s(\omega|Y, M))E([s(\omega|Y, M)']').
$$

(4.27)
In the following, we present the two approaches, showing how Louis’ formula is utilized in each method.

### 4.2.1 Recursively Approximated Standard Errors

Cai (2010a) proposed this method of utilizing Louis’ formula. The gist is that the information matrix of the observed log-likelihood is recursively approximated using the relationship between the observed information and the complete data log-likelihood. The MH-RM algorithm yields the elements of Equation 4.27 as by-products of iterations. The first and the second terms in Equation 4.27 can be approximated with Monte Carlo estimates as

\[
\hat{G}_{t+1} = \frac{1}{m_t} \sum_{j=1}^{m_t} \left[ H(\omega^{(t)}|Y, M^{(t+1)}_j) - s(\omega^{(t)}|Y, M^{(t+1)}_j)[s(\omega^{(t)}|Y, M^{(t+1)}_j)]' \right].
\] (4.28)

A recursive stochastic approximation may be used in order to reduce the noise of the estimate as follows:

\[
\hat{G}_{t+1} = \hat{G}_t + \epsilon_t \{ \hat{G}_{t+1} - \hat{G}_t \}. \] (4.29)

The last term in Equation 4.27 is the outer product of the complete data score vector in Equation 4.2, as suggested by Fisher’s identity in Equation 4.3. The score vector is approximated recursively in the same manner:

\[
\hat{s}_{t+1} = \hat{s}_t + \epsilon_t \{ \hat{s}_{t+1} - \hat{s}_t \}. \] (4.30)

With all the elements at hand, the observed data information matrix is approximated as

\[
I_{t+1} = \hat{G}_{t+1} + \hat{s}_{t+1}' \hat{s}_{t+1}'. \] (4.31)

While the recursive approximation approach has been known to perform better than the post-convergence approach, much research has pointed out that recursively approxi-
imated SEs may yield estimates that are not positive definite.

### 4.2.2 Post-convergence Standard Errors

This approach directly applies Louis (1982)'s equation, wherein Monte Carlo integration is used upon convergence of MH-RM (Diebolt & Ip, 1996). The convergence with respect to the ML estimate \( \hat{\omega} \) implies \( s(\omega|Y, M) = 0 \), and accordingly the last term in Equation 4.27 no longer exists when evaluated at \( \hat{\omega} \). Let \( m_c \) denote the number of samples drawn to approximate the information matrix. The first term in Equation 4.27 is approximated as

\[
E(H(\omega|Y, M)) \approx \frac{1}{m_c} \sum_{j=1}^{m_c} H(\hat{\omega}|Y, M_j),
\]

and similarly, the second term is computed as

\[
E(s(\omega|Y, M)[s(\omega|Y, M)]') \approx \frac{1}{m_c} \sum_{j=1}^{m_c} (s(\hat{\omega}|Y, M_j)[s(\hat{\omega}|Y, M_j)]').
\]

### 4.3 Scoring of Random Effects by Empirical Bayes

Inference of random effects/factor using the likelihood-based approach is known as the empirical Bayes (EB) prediction. Essentially, EB uses a prior distribution of random effects/factor, and inference is based on the conditional posterior predictive distribution of the missing data given ML estimates \( \hat{\omega} \) (McCulloch et al., 2008; Skrondal & Rabe-Hesketh, 2014). Specifically, an EB estimate of the missing data \( M \) is defined as \( E(M|Y; \hat{\omega}) \), which is derived from the following conditional posterior predictive distribution

\[
\Pi(M|Y; \hat{\omega}) = \frac{f(Y|M; \hat{\omega})\Pi(M)}{\int f(Y|M; \hat{\omega})\Pi(M)dM}.
\]

Scoring of random effects/factor by EB, i.e., approximating \( E(M|Y; \hat{\omega}) \), can be done with the same approaches as those used for estimating SEs. Hence by analogy, we will call each approach 1) recursively approximated random effects scoring and 2) post-
convergence random effects scoring. While our ultimate interest is in scoring of item random effects \( u_\alpha \) in particular, we present scoring for all random effects/factor with a generic notation \( M = \{ \theta, c_m, u_\alpha, u_\gamma \} \).

### 4.3.1 Recursively Approximated Random Effects Scores

For a Monte Carlo estimate of random effects/factor

\[
\mathbf{M}_{t+1} = \frac{1}{m_t} \sum_{j=1}^{m_t} \mathbf{M}_j^{(t+1)},
\]

(4.35)

their recursive SA is performed as

\[
\hat{\mathbf{M}}_{t+1} = \hat{\mathbf{M}}_t + \epsilon_t \{ \tilde{\mathbf{M}}_{t+1} - \hat{\mathbf{M}}_t \}.
\]

(4.36)

Upon convergence of MH-RM, empirical Bayes estimates of \( \hat{\mathbf{M}} = \{ \hat{\theta}, \hat{c}_m, \hat{u}_\alpha, \hat{u}_\gamma \} \) are obtained.

### 4.3.2 Post-convergence Random Effects Scores

Upon convergence, the ML solutions for the joint model are obtained by MH-RM. Given the ML estimates of model parameters \( \hat{\omega} \), EB prediction can be used for random effects/factor scoring with additional samples after convergence. Basically, \( \hat{\omega} \) is inserted to the model and the predicted means are estimated by averaging over sample means of the imputed random effects/factor as follows:

\[
E(\mathbf{M}|\mathbf{Y}; \hat{\omega}) \approx \frac{1}{m_c} \sum_{j=1}^{m_c} \mathbf{M}_j.
\]

(4.37)
4.4 Additional Notes on Estimation Strategies

This section discusses miscellaneous strategies that aid estimation. They include obtaining starting values for parameters in each of the three parts in the model along with some strategies for improving estimation and computation of SEs.

First, the starting values refer to the initial values that start the MH-RM cycles. These are different from the Stage III starting values, which are obtained by averaging the parameter estimates from Stage II after having run Stage I (see Section 4.1). They may be called “crude” as opposed to the “refined” starting values for Stage III (Monroe & Cai, 2014).

The parameters of the Part I model and the Part II model may be easily initialized with some “crude” values that are not based on any data. Very crude choices for item slopes and intercepts may be 1.0 and 0, respectively, and those for the mean and the standard deviation of the (standard) normal distribution may be 0 and 1. They could be certainly adjusted with any prior knowledge on or after some inspection of the data. However, as warned by Monroe (2014), in the case where the likelihood surface is irregular, it is crucial that we provide good starting values so that the Stage III starting values in the proximity of MLE are attained. Note that without stochastic imputations of missing data, and hence with noise-free functions, MH-RM is simply reduced to Newton’s method, or more specifically the Newton-Raphson method, which requires starting values to be sufficiently close to the solutions. Hence, we need to carefully choose the initial values for the parameters in the Part 3 model: the multivariate nonlinear cross-classified random effects model. This is very challenging in practice, and thus we may use a multistage method. It consists of the following three stages:

- **Stage 1**: Draw $\theta$ from the Part I density. The simulated value is $\theta^*$. 
- **Stage 2**: Draw $c_m$ from the Part II density conditional on $\theta = \theta^*$. The simulated value is $c_m^*$. 


• **Stage 3:** Estimate the parameters of fixed effects \( \omega_3 \) given \( c_{m}^{*} \).

As a quick simplifying measure, an approximate method can be applied in **Stage 2** and **Stage 3** to a subset of the data. The approximate methods include two types of approaches: linearization methods based on a Taylor series expansion (e.g., Lindstrom & Bates, 1990) and direct approximation or Laplace approximation (e.g., Wolfinger & Xihong, 1997). Note that the estimates need not be precise as long as they are reasonable to be used as the initial conditions for MH-RM. Another benefit of the multistage method is that it provides (as by-products) good starting values for the parameters in Part I and Part II as well.

Next, we provide some strategies necessary and/or helpful for implementing MH-RM for the proposed model. Considering the parameter spaces of the item slope and variance parameters, it is helpful to transform them to logarithmic scales to improve estimation. Furthermore, it is recommended to impose priors on the variance of random effects because estimating the variance can prove problematic in practice. Small variances of random effects parameters is a well-known issue in hierarchical models (e.g., Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013; Gelman et al., 2014; Schuurman, Grasman, & Hamaker, 2017). The situation is aggravated when the number of groups is small. In this study, the number of raters tends not to be large. In ML estimation, a prior is treated as a "soft constraint" or a penalty term added to the log-likelihood function. When a prior is imposed, the estimation becomes a maximum penalized likelihood (MPL) estimation. This strategy has been used in various contexts (e.g., Cai, Yang, & Hansen, 2011; Falk & Cai, 2016).

In applying the RM filter, a couple of strategies are necessary to stabilize estimation. First, as in Newton’s method, Hessian modification is applied in order to make positive definite Hessian matrix (Nocedal & Wright, 2006). In addition, as commonly used in stochastic optimization (Spall, 1998), “blocking” can be used to prevent “wild behaviors” of the objective function (see Monroe, 2014 for example). In essence, it skips an update of an estimate when the update does not satisfy a certain range-criterion (too large or
Lastly, as in the Part III model, when the likelihood surface is not regular, the fraction of missing information (FMI) can be high. This in turn may result in negative observation information matrix, in which case SEs cannot be computed. To resolve the issue, first we may rearrange the observed information matrix as follows:

\[
-\frac{\partial^2 l(\omega|Y)}{\partial \omega \partial \omega'} = E(H(\omega|Y, M)) - E(s(\omega|Y, M)[s(\omega|Y, M)]')
= (I - E(s(\omega|Y, M)[s(\omega|Y, M)]'))E(H(\omega|Y, M)^{-1}))E(H(\omega|Y, M)).
\] (4.38)

Observing Equation 4.38, we immediately recognize that the negative error variance is due to high FMI defined by the second term in the parentheses. Let the first term, \(I - E(s(\omega|Y, M)[s(\omega|Y, M)]')E(H(\omega|Y, M)^{-1}))\), be denoted \(R\). Then our goal is to make \(R\) positive definite via preconditioning by ridging (or eigenvalue modification). For a non-symmetric \(R\), singular value decomposition (SVD) can be implemented to enforce symmetry.

In addition, re-parameterizing the item slope parameters and random effect variances to be estimated in the log-scale requires a corresponding adjustment of the Fisher information matrix. For an arbitrary \(n\)-dimensional vector \(\omega\) to be re-parameterized to \(\log(\omega)\), the adjusted information for \(\omega\) must also be transformed as

\[
\mathcal{I} = J'\mathcal{I}(\log(\omega))J,
\] (4.39)

where \(J\) is an \(n \times n\) Jacobian matrix. The \((i,j)\)th element of \(J\) is

\[
J_{ij} = \frac{\partial \log(\omega_i)}{\partial \omega_j}.
\] (4.40)
CHAPTER 5

Application to English Language Proficiency Assessment

In this chapter, we apply the model and estimation scheme to the ELPA21 data. The usual practice in methods development is to first conduct and present simulation studies prior to real-data applications because this promotes the generality of a given method. Though our approach should be broadly applicable as well, the Braille ELP assessment from ELPA21 has been the main context throughout and is the root motivation for this study, and therefore the empirical application before simulation studies. In order to provide in-depth discussions, we take the whole data set that can be applied within the scope of this study rather than choosing one. Furthermore, by making comparisons to the standard approach that is commonly used in operation and the linear fixed-effects approach, we hope to show how our proposed approach yields differing conclusions and highlight the need to improve upon them. Lastly, the analysis of this empirical data set will serve as the basis for simulation designs in next chapter.

5.1 Data and Methods

The calibration data set of the original item pool is from the 2015-16 summative administration. The assessment has four domains (Listening, Reading, Writing, and Speaking) and consists of six gradebands (Grade K, Grade 1, Grade 2-3, Grade 4-5, Grade 6-8, and Grade 9-12). For convenience, we shall use “Grade K”, “Grade 1”, “Grade 3”, “Grade 5”, “Grade 7”, “Grade 9” for the corresponding gradebands. There are five or six test forms per gradeband that were randomly assigned to students. This planned missing design is aimed at maximizing the number of test items analyzed. The missing data by design
is treated as missing completely at random (MCAR; Rubin, 1987). Specifically, the test forms were constructed to have balanced task types (e.g., Classroom Table, Conversation, Picture Description, Opinion, Observe and Report for Grade 1 Speaking) based on the requirements of summative assessment blueprints. In addition, the test forms included tasks that do not contribute to students’ scores and only serve research purposes (English Language Proficiency for the 21st Century, 2017).

Table 5.1 presents the sample sizes and the number of items by each gradeband. Each sample size ranges from 37,305 to 55,602, and the total sample size is 306,070. There are 2,113 items total, and the distribution of items across domains and gradebands is also shown. The judgmental data were collected in the cut-score linking event held in 2018 as described in Chapter 1. Table 5.2 presents the information on the number of items and raters by domain and gradeband for modified Braille test items, which are of our interest. There are 3 to 5 raters but in most cases 4. Most items for Listening and Reading are dichotomous in nature while all of Speaking items are of polytomous nature. Also, it is shown that the Listening and Reading domains contain a plentiful number of modified Braille test items with original test item counterparts, whereas a substantial portion of Speaking items has been newly developed altogether.

Table 5.1: Data Summary: Online Items

<table>
<thead>
<tr>
<th>Gradeband</th>
<th>Sample Size</th>
<th>Number of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All Domains</td>
</tr>
<tr>
<td>K</td>
<td>37,305</td>
<td>429</td>
</tr>
<tr>
<td>1</td>
<td>37,705</td>
<td>381</td>
</tr>
<tr>
<td>3</td>
<td>70,984</td>
<td>350</td>
</tr>
<tr>
<td>5</td>
<td>55,189</td>
<td>343</td>
</tr>
<tr>
<td>7</td>
<td>55,602</td>
<td>291</td>
</tr>
<tr>
<td>9</td>
<td>49,285</td>
<td>319</td>
</tr>
<tr>
<td>Total</td>
<td>306,070</td>
<td>2,113</td>
</tr>
</tbody>
</table>
Table 5.2: Data Summary: Modified Braille Test Items

<table>
<thead>
<tr>
<th>Domain</th>
<th>Grade</th>
<th>Online</th>
<th>No Online</th>
<th>Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dichotomous</td>
<td>Polytomous</td>
<td>Total</td>
</tr>
<tr>
<td>Listening</td>
<td>K</td>
<td>19</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>28</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>23</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>25</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>57</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>30</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>Reading</td>
<td>K</td>
<td>21</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>34</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>39</td>
<td>2</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>42</td>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>44</td>
<td>6</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>53</td>
<td>3</td>
<td>56</td>
</tr>
<tr>
<td>Speaking</td>
<td>K</td>
<td>0</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Writing</td>
<td>K</td>
<td>9</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>6</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>3</td>
<td>15</td>
<td>18</td>
</tr>
</tbody>
</table>
As the scope of the work, for the time being, is dichotomous modified items, we selected the Listening and Reading domains for application. For analysis, we randomly selected 10,000 samples from responses from ELs from the general population on the original test items. For the judgmental data, we only used the first-round ratings. As a reminder, raters provided ratings at two cuts: cut 3 and cut 4. We used these two ratings per item.

We used an MH-RM algorithm coded and implemented in R by the author (R Core Team, 2018) to jointly estimate the unified model (Part I + Part II + Part III). For Stage III starting values, 3,000 Stage I iterations and 500 Stage II iterations were used with gain constants .1. At Stage III, the decreasing gain constants $\epsilon_t = \epsilon_0 / t^{0.75}$ were applied, where $\epsilon_0$ are the gain constants used in Stage I and II. A “burn-in” of 10 was used for an MH sampler, and simulation size $m_t$ was set to 1 across the stages. Convergence was determined by a window of 3 iterations within a $1.0 \times 10^{-4}$ tolerance threshold. After convergence, 25,000 Monte Carlo draws were used to score all random effects/factor. Post-convergence standard errors were computed from 250 draws with a thinning interval of 20. Further details is discussed in next chapter, Chapter 6.

5.2 Results

5.2.1 Descriptive Statistics

First, the descriptive statistics of modified Braille items in comparison to those of the original source items for cut 3 and 4 are shown in Table 5.3 and 5.4, respectively. It presents the mean, the standard deviation, the minimum, and the maximum of the probability of correct responses. For the Braille items, the quantities are calculated from judges’ ratings, i.e., expected probabilities of correct responses by BLV students, while for the source items, they are calculated from the actual proportion of correct responses from by the general EL population. Since there is a one-to-one correspondence between the modified Braille items and their source items, a direct comparison is possible. More
often than not, the mean expected probability of correct responses to the modified Braille test items judged by raters is slightly lower than that for the source items. In general, Grade 3 Listening, Grade 3 Reading, and Grade 5 Reading showed the greatest shift in the mean item difficulty from the original to Braille.

Table 5.3: Item Summary for Cut 3: Probability of Correct Responses

<table>
<thead>
<tr>
<th>Domain</th>
<th>Grade</th>
<th>Online</th>
<th>Braille</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean  SD  Min  Max</td>
<td>Mean  SD  Min  Max</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>0.57  0.15  0.33  0.80</td>
<td>0.53  0.15  0.29  0.79</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.68  0.14  0.36  0.93</td>
<td>0.67  0.14  0.35  0.93</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.59  0.16  0.28  0.82</td>
<td>0.47  0.18  0.20  0.77</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.56  0.18  0.16  0.84</td>
<td>0.43  0.18  0.10  0.75</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.58  0.16  0.26  0.91</td>
<td>0.52  0.16  0.21  0.82</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.55  0.19  0.24  0.92</td>
<td>0.50  0.19  0.20  0.91</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>0.54  0.14  0.28  0.82</td>
<td>0.52  0.15  0.24  0.80</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.57  0.23  0.22  0.97</td>
<td>0.56  0.23  0.19  0.96</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.74  0.22  0.18  1.00</td>
<td>0.62  0.18  0.32  0.92</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.52  0.22  0.09  0.90</td>
<td>0.40  0.24  0.04  0.83</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.65  0.19  0.30  0.96</td>
<td>0.60  0.20  0.24  0.93</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.53  0.19  0.20  0.98</td>
<td>0.50  0.20  0.15  0.95</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation; Min = Minimum; Max = Maximum

In addition, we present in Table 5.5 the proportion of items in each group of the relationship with source (the types of relationships between modified Braille items and their source were defined and grouped in Table 1.3). A large number of items belong to Group 1 and 2. The degree of modification in items and the resulting variation in expected item difficulty increases from Group 1 to 4. For example, in Group 1 are items that no other change is involved except for the bare-necessary modification in the format of presentation, and hence very little change in item difficulty is expectable. Items in Group 4, on the other hand, have different content that may result in no consistent relationship between item difficulties. However, we cannot assume that items in each group actually stick to the relationship with the original items in terms of item difficulty. This point will be further illustrated in Section 5.2.3.

Next, the agreement among the four raters for Listening and Reading is presented
Table 5.4: Item Summary for Cut 4: Probability of Correct Responses

<table>
<thead>
<tr>
<th>Domain</th>
<th>Grade</th>
<th>Online</th>
<th></th>
<th></th>
<th></th>
<th>Braille</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
<td>SD</td>
<td>Min</td>
</tr>
<tr>
<td>Listening</td>
<td>K</td>
<td>0.76</td>
<td>0.15</td>
<td>0.48</td>
<td>0.95</td>
<td>0.73</td>
<td>0.16</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.85</td>
<td>0.12</td>
<td>0.50</td>
<td>0.99</td>
<td>0.83</td>
<td>0.13</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.80</td>
<td>0.15</td>
<td>0.43</td>
<td>0.97</td>
<td>0.68</td>
<td>0.17</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.68</td>
<td>0.18</td>
<td>0.21</td>
<td>0.95</td>
<td>0.56</td>
<td>0.18</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.80</td>
<td>0.15</td>
<td>0.37</td>
<td>0.99</td>
<td>0.73</td>
<td>0.16</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.79</td>
<td>0.16</td>
<td>0.42</td>
<td>0.99</td>
<td>0.73</td>
<td>0.17</td>
<td>0.37</td>
</tr>
<tr>
<td>Reading</td>
<td>K</td>
<td>0.68</td>
<td>0.16</td>
<td>0.34</td>
<td>0.92</td>
<td>0.66</td>
<td>0.17</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.74</td>
<td>0.18</td>
<td>0.43</td>
<td>1.00</td>
<td>0.71</td>
<td>0.19</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.86</td>
<td>0.14</td>
<td>0.52</td>
<td>1.00</td>
<td>0.78</td>
<td>0.13</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.75</td>
<td>0.22</td>
<td>0.13</td>
<td>0.99</td>
<td>0.63</td>
<td>0.25</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.83</td>
<td>0.16</td>
<td>0.44</td>
<td>1.00</td>
<td>0.78</td>
<td>0.18</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.75</td>
<td>0.19</td>
<td>0.23</td>
<td>1.00</td>
<td>0.71</td>
<td>0.20</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation; Min = Minimum; Max = Maximum

Table 5.5: Item Summary: Relationship

<table>
<thead>
<tr>
<th>Domain</th>
<th>Grade</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>K</td>
<td>0.263</td>
<td>0.737</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.536</td>
<td>0.464</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.625</td>
<td>0.375</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.739</td>
<td>0.130</td>
<td>0.130</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.357</td>
<td>0.071</td>
<td>0.179</td>
<td>0.393</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.700</td>
<td>0.000</td>
<td>0.267</td>
<td>0.033</td>
</tr>
<tr>
<td>Reading</td>
<td>K</td>
<td>0.619</td>
<td>0.381</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.432</td>
<td>0.351</td>
<td>0.216</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.632</td>
<td>0.079</td>
<td>0.237</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.850</td>
<td>0.100</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.844</td>
<td>0.133</td>
<td>0.022</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.807</td>
<td>0.070</td>
<td>0.000</td>
<td>0.123</td>
</tr>
</tbody>
</table>
in Table 5.6 and 5.7, respectively, in terms of Pearson correlation. The tables show the aggregated correlation of cut 3 and 4 for each gradeband. By and large, the correlations are considerably high, ranging from .825 to .994 for Listening and from .599 to .991 for Reading. We can speculate that such high agreements among raters can be attributed to their knowledge of the proportions of correct response from the general EL population on the original items, each of which serves as the anchor for a modified Braille test item.

5.2.2 Model Parameter Estimates

What follows is the parameter estimates and standard errors. Since all that we need from the Part I model is the latent ability information, the model parameters of our interested are those of Part II and Part III. Note that eventually, we only need the intercept and the regression coefficient values from Part III to compute revised item parameters for the modified Braille test items. However, the mean and the variance of the modified item parameters in Part II and the variance components in Part III also provide us a sense of the distribution of the items and the raters. The results are compiled in Table 5.8.

From the small positive mean of the random item effects for the original source items in Part II, we find that the items are relatively easy overall. Note that $\mu_c$ is the mean of intercepts, which are obtained by the negative product of slope and difficulty parameter. Their variance ranges from .53345 to 7.28380, with considerably higher variability for Grade 3 Reading than any other domain/gradeband.

The Part III results show negative intercepts and regression coefficient close to 1 for all but Grade 3 Reading, which has positive intercept and regression coefficient less than 1. Either case suggests that the modified Braille test items tend to be more difficult than the original counterparts. This is expected from the descriptive statistics on the items (see Table 5.3 and 5.4). Also of interest is the variance of random item effects and rater effects. The small variance estimates overall are due to the small units of the probability scale. In addition, other than Grade 3 Reading, it can be seen that the variances of random residual item effects after accounting for the original item parameters are small.
Table 5.6: Rater Agreement: Listening

<table>
<thead>
<tr>
<th></th>
<th>Grade K</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
<td>R4</td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.945</td>
<td>0.943</td>
<td>0.955</td>
</tr>
<tr>
<td>R2</td>
<td>0.945</td>
<td>1.000</td>
<td>0.930</td>
<td>0.925</td>
</tr>
<tr>
<td>R3</td>
<td>0.943</td>
<td>0.930</td>
<td>1.000</td>
<td>0.952</td>
</tr>
<tr>
<td>R4</td>
<td>0.955</td>
<td>0.925</td>
<td>0.952</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Grade 1</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
<td>R4</td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.992</td>
<td>0.990</td>
<td>0.906</td>
</tr>
<tr>
<td>R2</td>
<td>0.992</td>
<td>1.000</td>
<td>0.983</td>
<td>0.930</td>
</tr>
<tr>
<td>R3</td>
<td>0.990</td>
<td>0.983</td>
<td>1.000</td>
<td>0.872</td>
</tr>
<tr>
<td>R4</td>
<td>0.906</td>
<td>0.930</td>
<td>0.872</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Grade 3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.825</td>
<td>0.880</td>
</tr>
<tr>
<td>R2</td>
<td>0.825</td>
<td>1.000</td>
<td>0.895</td>
</tr>
<tr>
<td>R3</td>
<td>0.880</td>
<td>0.895</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Grade 5</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.893</td>
<td>0.828</td>
</tr>
<tr>
<td>R2</td>
<td>0.893</td>
<td>1.000</td>
<td>0.925</td>
</tr>
<tr>
<td>R3</td>
<td>0.828</td>
<td>0.925</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Grade 7</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
<td>R4</td>
<td>R5</td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.991</td>
<td>0.982</td>
<td>0.920</td>
<td>0.886</td>
</tr>
<tr>
<td>R2</td>
<td>0.991</td>
<td>1.000</td>
<td>0.984</td>
<td>0.932</td>
<td>0.912</td>
</tr>
<tr>
<td>R3</td>
<td>0.982</td>
<td>0.984</td>
<td>1.000</td>
<td>0.919</td>
<td>0.867</td>
</tr>
<tr>
<td>R4</td>
<td>0.920</td>
<td>0.932</td>
<td>0.919</td>
<td>1.000</td>
<td>0.885</td>
</tr>
<tr>
<td>R5</td>
<td>0.886</td>
<td>0.912</td>
<td>0.867</td>
<td>0.885</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Grade 9</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
<td>R4</td>
<td>R5</td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.925</td>
<td>0.928</td>
<td>0.930</td>
<td>0.926</td>
</tr>
<tr>
<td>R2</td>
<td>0.925</td>
<td>1.000</td>
<td>0.993</td>
<td>0.992</td>
<td>0.981</td>
</tr>
<tr>
<td>R3</td>
<td>0.928</td>
<td>0.993</td>
<td>1.000</td>
<td>0.994</td>
<td>0.984</td>
</tr>
<tr>
<td>R4</td>
<td>0.930</td>
<td>0.992</td>
<td>0.994</td>
<td>1.000</td>
<td>0.988</td>
</tr>
<tr>
<td>R5</td>
<td>0.926</td>
<td>0.981</td>
<td>0.984</td>
<td>0.988</td>
<td>1.000</td>
</tr>
<tr>
<td>Grade</td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
<td>R4</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>Grade K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.979</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.988</td>
<td>0.991</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.985</td>
<td>0.982</td>
<td>0.990</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.973</td>
<td>0.971</td>
<td>0.972</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.973</td>
<td>1.000</td>
<td>0.976</td>
<td>0.978</td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.971</td>
<td>0.976</td>
<td>1.000</td>
<td>0.972</td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.972</td>
<td>0.978</td>
<td>0.972</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.867</td>
<td>0.599</td>
<td>0.911</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.867</td>
<td>1.000</td>
<td>0.658</td>
<td>0.908</td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.599</td>
<td>0.658</td>
<td>1.000</td>
<td>0.664</td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.911</td>
<td>0.908</td>
<td>0.664</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Grade 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.940</td>
<td>0.893</td>
<td>0.949</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.940</td>
<td>1.000</td>
<td>0.894</td>
<td>0.936</td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.893</td>
<td>0.894</td>
<td>1.000</td>
<td>0.933</td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.949</td>
<td>0.936</td>
<td>0.933</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Grade 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.974</td>
<td>0.957</td>
<td>0.970</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.974</td>
<td>1.000</td>
<td>0.979</td>
<td>0.991</td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.957</td>
<td>0.979</td>
<td>1.000</td>
<td>0.974</td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.970</td>
<td>0.991</td>
<td>0.974</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Grade 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>1.000</td>
<td>0.956</td>
<td>0.946</td>
<td>0.976</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.956</td>
<td>1.000</td>
<td>0.948</td>
<td>0.958</td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.946</td>
<td>0.948</td>
<td>1.000</td>
<td>0.950</td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.976</td>
<td>0.958</td>
<td>0.950</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>
(from .00014 to .06373). Clearly, the original item difficulty parameters may have caused a huge reduction in item variability of the modified Braille items, and this implies that there is a substantially consistent relationship with the original item difficulties. It is also not a surprise to see the small variances of random residual rater effects as this was suggested by high rater agreement (correlation).
<table>
<thead>
<tr>
<th>Domain</th>
<th>Grade</th>
<th>Part II</th>
<th>Part III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>µ</td>
<td>σ²</td>
<td>µ</td>
</tr>
<tr>
<td>Listening</td>
<td>K</td>
<td>0.73816 0.53345</td>
<td>-0.23592 1.04232</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.16729) (0.17315)</td>
<td>(0.04689) (0.05028) (0.00398) (0.05068) (0.00045)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.02460 0.94651</td>
<td>-0.07494 1.01878</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.17942) (0.25857)</td>
<td>(0.00012) (0.00005) (0.00000) (0.00035) (0.00026)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.81866 0.97784</td>
<td>-0.68437 1.10777</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.20093) (0.28244)</td>
<td>(0.25804) (0.09712) (0.01674) (0.03315) (0.00118)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.68685 1.47547</td>
<td>-0.50493 0.98319</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.25504) (0.43726)</td>
<td>(0.12496) (0.05484) (0.01130) (0.00014) (0.00098)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.66366 1.18355</td>
<td>-0.32421 1.06199</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.20507) (0.31775)</td>
<td>(0.05677) (0.04338) (0.00685) (0.02471) (0.00056)</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>1.31879 1.36448</td>
<td>-0.18640 1.03107</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.21343) (0.35209)</td>
<td>(0.00135) (0.00078) (0.00002) (0.00283) (0.00023)</td>
</tr>
<tr>
<td>Reading</td>
<td>K</td>
<td>0.47203 0.48184</td>
<td>-0.10010 1.00176</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.15087) (0.14870)</td>
<td>(0.02464) (0.03376) (0.00096) (0.01141) (0.00003)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1.08353 2.37420</td>
<td>-0.07176 0.97599</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.25196) (0.55227)</td>
<td>(0.11954) (0.05656) (0.00453) (0.01070) (0.00012)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.40357 7.28380</td>
<td>0.17607 0.31779</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.43664) (1.68095)</td>
<td>(0.16632) (0.03884) (0.06584) (0.03252) (0.00147)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.81514 1.75305</td>
<td>-0.62341 1.07968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.20562) (0.39909)</td>
<td>(0.09832) (0.05574) (0.00845) (0.00005) (0.00058)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.07762 1.24241</td>
<td>-0.16273 1.03942</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.16678) (0.25595)</td>
<td>(0.02806) (0.01922) (0.00222) (0.00114) (0.00027)</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.35702 1.10862</td>
<td>-0.07178 1.04753</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.13863) (0.20656)</td>
<td>(0.04921) (0.03508) (0.00558) (0.00303) (0.00036)</td>
</tr>
</tbody>
</table>

Note. Standard errors are in parentheses. We present the values up to five decimal points due to the small magnitudes of the variance estimates, which result from small units of the probability scale.
5.2.3 Revised Item Parameters

We now examine how well the distribution of item parameters is reproduced by plotting the predicted probability against the observed probability on the modified Braille test items. The observed probability is simply the average of all expected probabilities of correct responses assessed by the raters. Accordingly, it also represents the predicted probability by the linear fixed-effect approach. The probability of having a correct response for item $i$ at cut $k$ predicted with the multivariate nonlinear random-effect modeling is computed as $1/(1 + \exp[a_i^* \theta_k^* + \tilde{c}_i])$, where $\tilde{c}_i$ is the revised item difficulty parameter estimate for item $i$. Recall that $\tilde{c}_i = \hat{\mu} + \hat{\pi}_c_i + \hat{\zeta}_i$. Figure 5.1 and 5.2 present plots showing the relationship between the observed probability and the predicted probability for Listening and Reading, respectively. They are based on the aggregated data at both cut 3 and cut 4. They are each denoted as the lower cut ($\theta_L^*$) and the higher cut ($\theta_H^*$).

The well-aligned plots indicate that the distribution of the item difficulty parameters for the modified Braille test items are well reconstructed to reflect judgments provided by the raters. On the other hand, the residual variations actually suggest that our random-effect approach would provide improvements over the linear fixed-effect approach. This is more prominent in Reading domain, Grade 3 Reading, Grade 9 Reading and Grade 9 Reading in particular. Also, the alignment is quite off for a few items in Grade 3 Listening.

Next, we investigate the item characteristic curves (ICCs) of the proposed random-effect approach and fixed-effect approach. As an illustration, we present the ICCs of Grade 3 Reading. We selected this gradeband/domain because it exhibited one of the largest difference in the probability of correct responses (see Table 5.3 and 5.4) and also low rater agreement (see Table 5.7). Also, we would find salient differences between the fixed-effects approach and our random-effects approach (see Figure 5.2). Importantly, it consists of items in every group (Group 1-4) in terms of the relationship with the original items (see Table 5.5). Figure 5.3 - 5.9 present plots depicting the difference between the ICCs of the fixed-effect and the random-effect approaches, respectively denoted 'Braille
The ICCs of the original items are also presented with markers indicating the proportion of correct responses at $\theta^*_{L}$ and $\theta^*_{H}$. In addition, the ICCs of the modified Braille items based on four raters are shown. Note that the two approaches yield different item parameters, and hence ICCs. Overall, for Grade 3 Reading, most of the modified Braille items are expected to be more difficult than their original counterparts. Let us closely examine each group.

Figure 5.3 - 5.5 show the items in Group 1. While little changes in item parameters are expected for a large number of items, there exist some (e.g., Item 4, Item 20, Item 22, Item 24, Item 30) that have non-negligible shifts, inflicting more than 10% decrease in probability at the two cuts. The probability of correct response for Item 22, in particular, decreased by 20% at lower cut (see Figure 5.5). It is clear that we should not assume the same item parameters as the original items even for Group 1 items. These individual item plots also show that we may score differently with our random-effects approach from the linear fixed-effects approach, as previously suggested by their comparison in Figure 5.2. For Item 9, for example, the expected probability at cut 4 is particularly different for two approaches. Another interesting example is Item 11. The revised item parameter by our random-effects approach would make a remarkable difference for students with low abilities, even though the expected probabilities from two approaches are not so different at the two cuts (see Figure 5.4). Figure 5.6 shows the items in Group 2. Like those in Group 1, the degree of change in item difficulties varies across items. For Item 25 and 26, only a small decrease in probability is expected, whereas there is a huge increase for Item 27.

Figure 5.7 - 5.8 show the items in Group 3. These items are distinguishable from those in other groups in that the slopes are extremely high. Consequently, the ICCs obtained from our random-effects approach is very much different from the fixed-effects approach. For a comparison with the original version and the fixed-effects approach, take Item 34 for an example. A student with a latent ability score slightly above -2 is likely to get this item correct when taking the original version, while it is highly un-
likely on the Braille version. If the fixed-effect approach is applied, then there is about 50% chance of getting it correctly on the Braille version. Additionally, it is worthwhile to note that the current cut-scores may not be optimal given that we do not see huge differences in the probability of correct responses at the two cut-scores. This was also observed in the above example (Item 11 in Group 1). We will elaborate on this in Chapter 7. Lastly, Figure 5.9 shows the items in Group 4. As a reminder, these modified items have different contents entirely from their source items, so we cannot expect any consistent relationship between item difficulties. Item 42 showed quite a difference in item difficulty between the original and the Braille version, while Item 41 did not.

Finally, we constructed a hypothetical scoring table in Table 5.9. It is basically a ‘summed-score-to-scale-score conversion table’. Based on 10,000 responses from ELs from the general population on 38 modified items, the expected a posteriori (EAP) scores were obtained with the original item parameters. Then, with the same data set, EAP was scored again with revised item parameters from the three-part model. As a comparison, EAP from the item parameters from the linear fixed-effects approach was also obtained. Each summed score (0 – 38) has a corresponding EAP score. In addition, we converted the EAP scores into scale scores reported in ELPA21, which are also presented\(^1\).

The EAP scores or the scale scores in ELPA21 from the original item parameters and the revised item parameters markedly differ, with those from the revised item parameters being higher. It indicates that, had they not been scored fairly with the revised item parameters, BLV students would receive lower scores than what they deserve, which may lead to a substantively different decision (e.g. for their placement). It is also noteworthy to see how the two different approaches to revised item parameters yield differing EAP or the scale scores. To illustrate this, suppose the lower cut is - 0.8. A student who correctly responded to 15 items out of 38 would be classified as Level 3. However, he/she would be classified as Level 2 if the original parameters were applied. The linear fixed-effects approach would also misclassify him/her as Level 2.

\(^1\)The scale score for Reading is defined as \(A\theta_{EAP} + D\) where \(A = 80\) and \(D = 550\).
We created an additional hypothetical scoring table by building a hypothetical test form based on the Group 1 items only. The purpose of this additional analysis is to see how reasonable it is to presume that item parameters change for Group 1 if we allow them to change based on experts’ judgments. This amounts to empirically testing the assumption made for Group 1 items (see Section 1.1.2) and hence examining whether it is appropriate to use the standard approach. Table 5.10 displays the scoring table for the test form consisting of Group 1 items, based on the same aforementioned procedure. The possible summed score from the 24 Group 1 items is 0 – 24.

The differences in the EAP or the scale scores in ELPA21 from the original item parameters and the revised item parameters are somewhat reduced from when the whole items are used (see Table 5.9), but meaningful differences remain. Again with the lower cut - 0.8, for example, a student who correctly responded to 9 items out of 24 would be classified as Level 3 with the revised item parameters (with either approach). However, he/she would be classified as Level 2 if the original parameters were applied. Clearly for the modified Braille test form, we cannot simply use the item parameters obtained from the original calibration.
Figure 5.1: Observed Probability vs. Expected Probability: Listening
Figure 5.2: Observed Probability vs. Expected Probability: Reading
Figure 5.3: Item Characteristic Curves for Group 1: Grade 3 Reading (Items 1 - 8)
Figure 5.4: Item Characteristic Curves for Group 1: Grade 3 Reading (Items 9 - 16)
Figure 5.5: Item Characteristic Curves for Group 1: Grade 3 Reading (Items 17 - 24)
Figure 5.6: Item Characteristic Curves for Group 2: Grade 3 Reading (Items 1 - 3)
Figure 5.7: Item Characteristic Curves for Group 3: Grade 3 Reading (Items 1 - 8)
Figure 5.8: Item Characteristic Curves for Group 3: Grade 3 Reading (Item 9)
Figure 5.9: Item Characteristic Curves for Group 4: Grade 3 Reading (Items 1 - 2)
Table 5.9: Scoring Table: Grade 3 Reading

<table>
<thead>
<tr>
<th>Summed Score</th>
<th>EAP</th>
<th>Original</th>
<th>Revised (R)</th>
<th>Revised (F)</th>
<th>Scale Score</th>
<th>Original</th>
<th>Revised (R)</th>
<th>Revised (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.962</td>
<td>-2.402</td>
<td>-3.115</td>
<td>313</td>
<td>358</td>
<td>301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-2.775</td>
<td>-2.173</td>
<td>-2.916</td>
<td>328</td>
<td>376</td>
<td>317</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-2.615</td>
<td>-1.977</td>
<td>-2.722</td>
<td>341</td>
<td>392</td>
<td>332</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-2.475</td>
<td>-1.808</td>
<td>-2.537</td>
<td>352</td>
<td>405</td>
<td>347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-2.353</td>
<td>-1.659</td>
<td>-2.363</td>
<td>362</td>
<td>417</td>
<td>361</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-2.245</td>
<td>-1.528</td>
<td>-2.199</td>
<td>370</td>
<td>428</td>
<td>374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-2.147</td>
<td>-1.410</td>
<td>-2.044</td>
<td>378</td>
<td>437</td>
<td>386</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-2.057</td>
<td>-1.303</td>
<td>-1.898</td>
<td>385</td>
<td>446</td>
<td>398</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-1.972</td>
<td>-1.206</td>
<td>-1.758</td>
<td>392</td>
<td>454</td>
<td>409</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-1.892</td>
<td>-1.116</td>
<td>-1.626</td>
<td>399</td>
<td>461</td>
<td>420</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-1.813</td>
<td>-1.033</td>
<td>-1.498</td>
<td>405</td>
<td>467</td>
<td>430</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-1.736</td>
<td>-0.954</td>
<td>-1.375</td>
<td>411</td>
<td>474</td>
<td>440</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-1.658</td>
<td>-0.880</td>
<td>-1.256</td>
<td>417</td>
<td>480</td>
<td>450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-1.579</td>
<td>-0.809</td>
<td>-1.140</td>
<td>424</td>
<td>485</td>
<td>459</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-1.497</td>
<td>-0.741</td>
<td>-1.027</td>
<td>430</td>
<td>491</td>
<td>468</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-1.412</td>
<td>-0.674</td>
<td>-0.916</td>
<td>437</td>
<td>496</td>
<td>477</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-1.324</td>
<td>-0.607</td>
<td>-0.806</td>
<td>444</td>
<td>501</td>
<td>486</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-1.233</td>
<td>-0.541</td>
<td>-0.697</td>
<td>451</td>
<td>507</td>
<td>494</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-1.138</td>
<td>-0.475</td>
<td>-0.589</td>
<td>459</td>
<td>512</td>
<td>503</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-1.040</td>
<td>-0.406</td>
<td>-0.481</td>
<td>467</td>
<td>518</td>
<td>512</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>-0.938</td>
<td>-0.335</td>
<td>-0.372</td>
<td>475</td>
<td>523</td>
<td>520</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>-0.831</td>
<td>-0.261</td>
<td>-0.262</td>
<td>484</td>
<td>529</td>
<td>529</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>-0.720</td>
<td>-0.182</td>
<td>-0.151</td>
<td>492</td>
<td>535</td>
<td>538</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>-0.604</td>
<td>-0.098</td>
<td>-0.038</td>
<td>502</td>
<td>542</td>
<td>547</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>-0.483</td>
<td>-0.009</td>
<td>0.078</td>
<td>511</td>
<td>549</td>
<td>556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>-0.357</td>
<td>0.087</td>
<td>0.197</td>
<td>521</td>
<td>557</td>
<td>566</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>-0.225</td>
<td>0.191</td>
<td>0.321</td>
<td>532</td>
<td>565</td>
<td>576</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>-0.087</td>
<td>0.303</td>
<td>0.449</td>
<td>543</td>
<td>574</td>
<td>586</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.058</td>
<td>0.424</td>
<td>0.583</td>
<td>555</td>
<td>584</td>
<td>597</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.211</td>
<td>0.555</td>
<td>0.724</td>
<td>567</td>
<td>594</td>
<td>608</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.374</td>
<td>0.699</td>
<td>0.873</td>
<td>580</td>
<td>606</td>
<td>620</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.548</td>
<td>0.857</td>
<td>1.033</td>
<td>594</td>
<td>619</td>
<td>633</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.737</td>
<td>1.030</td>
<td>1.206</td>
<td>609</td>
<td>632</td>
<td>646</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.944</td>
<td>1.223</td>
<td>1.394</td>
<td>626</td>
<td>648</td>
<td>662</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1.174</td>
<td>1.437</td>
<td>1.601</td>
<td>644</td>
<td>665</td>
<td>678</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>1.430</td>
<td>1.677</td>
<td>1.831</td>
<td>664</td>
<td>684</td>
<td>696</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>1.718</td>
<td>1.948</td>
<td>2.087</td>
<td>687</td>
<td>706</td>
<td>717</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>2.045</td>
<td>2.254</td>
<td>2.375</td>
<td>714</td>
<td>730</td>
<td>740</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. "Original" refers to the original item parameters. "Revised (R)" refers to the revised item parameters obtained from our random-effects approach. "Revised (F)" refers to the revised item parameters obtained from the linear fixed-effects approach.
Table 5.10: Scoring Table: Grade 3 Reading (Group 1 Only)

<table>
<thead>
<tr>
<th>Summed Score</th>
<th>EAP</th>
<th>Scale Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Revised (R)</td>
</tr>
<tr>
<td>0</td>
<td>-2.758</td>
<td>-2.475</td>
</tr>
<tr>
<td>1</td>
<td>-2.475</td>
<td>-2.179</td>
</tr>
<tr>
<td>2</td>
<td>-2.228</td>
<td>-1.924</td>
</tr>
<tr>
<td>3</td>
<td>-2.008</td>
<td>-1.701</td>
</tr>
<tr>
<td>4</td>
<td>-1.808</td>
<td>-1.502</td>
</tr>
<tr>
<td>5</td>
<td>-1.626</td>
<td>-1.321</td>
</tr>
<tr>
<td>6</td>
<td>-1.455</td>
<td>-1.155</td>
</tr>
<tr>
<td>7</td>
<td>-1.295</td>
<td>-0.999</td>
</tr>
<tr>
<td>8</td>
<td>-1.140</td>
<td>-0.850</td>
</tr>
<tr>
<td>9</td>
<td>-0.990</td>
<td>-0.706</td>
</tr>
<tr>
<td>10</td>
<td>-0.842</td>
<td>-0.565</td>
</tr>
<tr>
<td>11</td>
<td>-0.696</td>
<td>-0.425</td>
</tr>
<tr>
<td>12</td>
<td>-0.549</td>
<td>-0.283</td>
</tr>
<tr>
<td>13</td>
<td>-0.400</td>
<td>-0.140</td>
</tr>
<tr>
<td>14</td>
<td>-0.247</td>
<td>0.007</td>
</tr>
<tr>
<td>15</td>
<td>-0.090</td>
<td>0.160</td>
</tr>
<tr>
<td>16</td>
<td>0.073</td>
<td>0.319</td>
</tr>
<tr>
<td>17</td>
<td>0.244</td>
<td>0.487</td>
</tr>
<tr>
<td>18</td>
<td>0.426</td>
<td>0.666</td>
</tr>
<tr>
<td>19</td>
<td>0.623</td>
<td>0.858</td>
</tr>
<tr>
<td>20</td>
<td>0.837</td>
<td>1.067</td>
</tr>
<tr>
<td>21</td>
<td>1.074</td>
<td>1.296</td>
</tr>
<tr>
<td>22</td>
<td>1.339</td>
<td>1.549</td>
</tr>
<tr>
<td>23</td>
<td>1.637</td>
<td>1.834</td>
</tr>
<tr>
<td>24</td>
<td>1.976</td>
<td>2.154</td>
</tr>
</tbody>
</table>

*Note.* “Original” refers to the original item parameters. “Revised (R)” refers to the revised item parameters obtained from our random-effects approach. “Revised (F)” refers to the revised item parameters obtained from the linear fixed-effects approach.
A Monte Carlo simulation study was conducted in order to verify the proposed modeling for moderated item calibration. Parameter recovery and the accuracy of standard errors are typically evaluated in a simulation study. Notably, in this study, the recovery of revised item parameters for the modified test items are also of interest. In the end, this simulation study will provide guidance for future data collection and study design.

6.1 Simulation Study Design

6.1.1 Data Generation

We selected simulation conditions such that they both reflect and supplement the real settings of ELPA21. Since the primary interest is in the judgmental data and its model, we mostly manipulated conditions in the Part III model. In other words, we considered the Part I and II models as givens and fixed the parameters across all conditions.

Part I. The data were generated using a unidimensional 2PL IRT model for \( J = 20 \) unmodified items, with sample size \( P = 1000 \). True latent ability scores \( \theta \) were generated from a standard normal distribution. The slope parameters were sampled from a normal (1.672, .761) distribution, and the intercepts were sampled from a normal (2.308, 1.670) distribution. The slope parameters were truncated at .5. The intercepts were also constrained such that item difficulty, defined as the negative ratio of the intercepts to slope parameters, is between -3 and 3. These were chosen to reflect the properties of the unmodified item pool in the ELP assessment.
Part II. The data were generated using a cross-classified random effects 2PL IRT model for \( I = 20 \) and 50 dichotomous items with sample size \( P = 1000 \). The slope parameters were drawn from normal \((1.260, 0.534)\) distribution, truncated at .5. Note that the slope parameters are of interest and thus fixed to the true values in the analysis. The intercept parameters follow normal \((1.287, 1.230)\) distribution. It should be noted that actual intercept parameters differ for each replication because items were treated as random. Here, the conditions mirror the properties of the modified item pool in the ELP assessment.

Part III. Following ELPA21, we only consider two cut-scores \((K = 2)\). The two cut-scores were set at \( \theta^* = (-0.5, 0.2)' \). We fixed the intercept and regression coefficients for item covariate \( c_i \) as \( \mu = 0 \) and \( \pi = .6 \) and the error covariance matrix is specified as

\[
\Sigma_e = \begin{bmatrix}
.001 & .002 \\
.001 & .002
\end{bmatrix}.
\] (6.1)

Manipulated factors are the number of raters, item variance and rater variance (the magnitude of discrepancy in rater judgments) after accounting for the item covariate. For \( I = 20 \) items, The number of raters was \( R = 2, 3, 4, \) and 10, the item residual variance \( \sigma^2_\zeta = (0.2^2, 0.5^2) \), and the rater variance \( \sigma^2_\gamma = (0.1^2, 0.2^2) \). Four additional conditions were considered; for \( R = 2 \), we increased the number of items to \( I = 50 \) (as in Part II), all other factors being the same. This aims to see how an increased number of items more precisely estimates the parameters in the Part II model, which is important since those estimates may impact the accuracy of Part III model parameter estimates. The simulation conditions are presented in Table 6.1.

Accordingly, a total of 20 conditions (= 4 number of raters \( \times \) 2 item variance \( \times \) 2 rater variance + 4 additional conditions) were used to generate the rating data from the multivariate nonlinear cross-classified random effects model (Equation 3.10). First, we generated “true” revised item parameter \( \tilde{c}_i \) by simply adding random item effects. That is, we used a model with no rater effect. This simulates an ideal scenario where an
Table 6.1: Simulation Condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>I</th>
<th>R</th>
<th>$\sigma_\xi^2$</th>
<th>$\sigma_\gamma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>2</td>
<td>0.04</td>
<td>0.009</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>2</td>
<td>0.25</td>
<td>0.009</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>2</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>3</td>
<td>0.04</td>
<td>0.009</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>3</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>3</td>
<td>0.25</td>
<td>0.009</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>3</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>4</td>
<td>0.04</td>
<td>0.009</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>4</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>4</td>
<td>0.25</td>
<td>0.009</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>4</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>10</td>
<td>0.04</td>
<td>0.009</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>10</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>10</td>
<td>0.25</td>
<td>0.009</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>10</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>2</td>
<td>0.04</td>
<td>0.009</td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>2</td>
<td>0.25</td>
<td>0.009</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>2</td>
<td>0.25</td>
<td>0.04</td>
</tr>
</tbody>
</table>

infinite pool of raters come to a perfect agreement. These parameters are therefore the target to recover, which is an ultimate goal of this study. Again, “true” revised item parameters differ for each replication because items are treated as random. Then, rater-variations were added to reflect raters’ judgements. The probability scale of the ratings constrain the data to be between 0 and 1.

For each condition, 100 data sets were generated. All data generation was performed in R program (R Core Team, 2018).

6.1.2 Analysis

We used the MH-RM algorithm as implemented in R by the author (R Core Team, 2018) to jointly estimate the unified model (Part I + Part II + Part III). To obtain a starting value, we used a multistage estimation method as described in Section 4.4. Specifically, we used
flexMIRT (Cai, 2015) for the Part I model, the Laplace approximations (Wolfinger, 1993) as implemented in SAS NL MIXED for the Part II model, the cross-classified random effects IRT model and a linearization-based method (Lindstrom & Bates, 1990) available on SAS %NLINMIX Macro for the Part III model, the multivariate nonlinear cross-classified random effects model (SAS Institute Inc., 2013).

For the Stage III starting values, 300 Stage I iterations and 700 Stage II iterations were used with gain constants .1 for the Part I and II models and \( \epsilon_t = (.05, .05, .05, .05, .1, .1, .1)' \) for the Part III model. As pointed out by Cai (2008), vector-valued gain constants are allowed to accommodate differential rate of convergence due to the difference in the fraction of missing information, and they can be determined by monitoring changes in estimates across iterations (see Kesten, 1958). At Stage III, decreasing gain constants \( \epsilon_t = \epsilon_0 / t^{0.75} \) were applied, where \( \epsilon_0 \) are the gain constants used in Stage I and II. After monitoring the traces of parameter estimates, a “burn-in” of 10 was sufficient for an MH sampler when combined with the RM update. The simulation size \( m_t \) was set to 1 across stages. Convergence was determined by examining a window of 3 iterations with a \( 1.0 \times 10^{-4} \) tolerance. After convergence, 25,000 Monte Carlo draws were used to score all random effects/factor. Post-convergence standard errors were computed from 250 draws with thinning interval of 100.

### 6.1.3 Evaluation Criteria

First, we assessed parameter recovery with estimated bias and root mean squared error (RMSE). Given a parameter \( \omega \) and its MLE \( \hat{\omega}_n \) in replication \( n \), the estimated bias for \( \omega \) is defined as

\[
\text{Bias}_{\omega} = N^{-1} \sum_{n=1}^{N} (\hat{\omega}_n - \omega),
\]  

(6.2)

where \( N \) is number of Monte Carlo replications. The RMSE for \( \omega \) is defined as

\[
\text{RMSE}_{\omega} = \sqrt{N^{-1} \sum_{n=1}^{N} (\hat{\omega}_n - \omega)^2}.
\]  

(6.3)
Second, we measured the accuracy of standard errors by comparing the average of estimated post-convergence standard errors against the Monte Carlo standard deviation of point estimates. Let \( se(\hat{\omega}_n) \) be the estimated standard error of \( \omega \) in replication \( n \). Then, each is computed as

\[
E[se(\hat{\omega})] = N^{-1} \sum_{n=1}^{N} se(\hat{\omega}_n),
\]

and

\[
SD(\hat{\omega}) = \sqrt{\frac{(N - 1)^{-1} \sum_{n=1}^{N} (\hat{\omega}_n - \bar{\omega})^2},}
\]

where \( \bar{\omega} \) is the mean of the point estimates for all replications.

Lastly, the recovery of random item effects \( u_\alpha \), which is the revised item parameter, is particularly important in this study. Recall that the revised item parameter is defined as \( \bar{\alpha}_i = \mu + \pi c_i + \zeta_i \). We used Person correlations, bias and RMSE to examine the recovery of averaged \( u_\alpha \).

### 6.2 Simulation Study Results

Simulation study results will be presented in the following order: the recovery of parameters, the accuracy of standard errors, and the recovery of random item effects that correspond to the defined revised item parameters.

#### 6.2.1 Point Estimates

Since the item parameters of the Part I model were fixed across conditions, it suffices for us to present the result from just one condition. Though the estimates are not exactly the same across conditions, the differences are quite minimal. The generating parameter values and parameter estimates, estimated bias, and RMSE for the Part I model from Condition 1 are summarized in Table 6.2. It shows quality recovery of the slopes.
and intercepts. It appears that relatively high bias and RMSE are associated with high generating values of slopes.

Table 6.2: Estimates of Part I

<table>
<thead>
<tr>
<th>Item</th>
<th>Intercepts True</th>
<th>Estimates</th>
<th>Bias</th>
<th>RMSE</th>
<th>Slopes Item</th>
<th>True Estimates</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.122</td>
<td>1.163</td>
<td>0.041</td>
<td>0.110</td>
<td>1</td>
<td>1.676</td>
<td>0.026</td>
<td>0.130</td>
</tr>
<tr>
<td>2</td>
<td>2.737</td>
<td>2.793</td>
<td>0.056</td>
<td>0.194</td>
<td>2</td>
<td>1.965</td>
<td>0.020</td>
<td>0.195</td>
</tr>
<tr>
<td>3</td>
<td>1.896</td>
<td>1.915</td>
<td>0.019</td>
<td>0.110</td>
<td>3</td>
<td>1.390</td>
<td>-0.010</td>
<td>0.132</td>
</tr>
<tr>
<td>4</td>
<td>1.728</td>
<td>1.762</td>
<td>0.034</td>
<td>0.153</td>
<td>4</td>
<td>2.162</td>
<td>0.019</td>
<td>0.197</td>
</tr>
<tr>
<td>5</td>
<td>0.719</td>
<td>0.730</td>
<td>0.011</td>
<td>0.073</td>
<td>5</td>
<td>1.504</td>
<td>1.489</td>
<td>-0.015</td>
</tr>
<tr>
<td>6</td>
<td>2.233</td>
<td>2.261</td>
<td>0.028</td>
<td>0.149</td>
<td>6</td>
<td>1.924</td>
<td>1.922</td>
<td>-0.002</td>
</tr>
<tr>
<td>7</td>
<td>0.997</td>
<td>1.039</td>
<td>0.042</td>
<td>0.118</td>
<td>7</td>
<td>2.507</td>
<td>2.519</td>
<td>0.012</td>
</tr>
<tr>
<td>8</td>
<td>-0.477</td>
<td>-0.460</td>
<td>0.017</td>
<td>0.082</td>
<td>8</td>
<td>2.003</td>
<td>2.023</td>
<td>0.020</td>
</tr>
<tr>
<td>9</td>
<td>1.673</td>
<td>1.704</td>
<td>0.031</td>
<td>0.106</td>
<td>9</td>
<td>1.424</td>
<td>1.435</td>
<td>0.011</td>
</tr>
<tr>
<td>10</td>
<td>3.843</td>
<td>3.874</td>
<td>0.031</td>
<td>0.240</td>
<td>10</td>
<td>2.546</td>
<td>2.553</td>
<td>0.007</td>
</tr>
<tr>
<td>11</td>
<td>1.347</td>
<td>1.365</td>
<td>0.018</td>
<td>0.116</td>
<td>11</td>
<td>2.428</td>
<td>2.438</td>
<td>0.010</td>
</tr>
<tr>
<td>12</td>
<td>3.323</td>
<td>3.369</td>
<td>0.046</td>
<td>0.225</td>
<td>12</td>
<td>2.089</td>
<td>2.089</td>
<td>0.000</td>
</tr>
<tr>
<td>13</td>
<td>-0.394</td>
<td>-0.373</td>
<td>0.021</td>
<td>0.071</td>
<td>13</td>
<td>1.854</td>
<td>1.863</td>
<td>0.009</td>
</tr>
<tr>
<td>14</td>
<td>2.215</td>
<td>2.224</td>
<td>0.009</td>
<td>0.139</td>
<td>14</td>
<td>1.194</td>
<td>1.204</td>
<td>0.010</td>
</tr>
<tr>
<td>15</td>
<td>3.175</td>
<td>3.232</td>
<td>0.057</td>
<td>0.252</td>
<td>15</td>
<td>2.707</td>
<td>2.764</td>
<td>0.057</td>
</tr>
<tr>
<td>16</td>
<td>2.811</td>
<td>2.783</td>
<td>-0.028</td>
<td>0.184</td>
<td>16</td>
<td>1.215</td>
<td>1.179</td>
<td>-0.036</td>
</tr>
<tr>
<td>17</td>
<td>2.484</td>
<td>2.590</td>
<td>0.106</td>
<td>0.243</td>
<td>17</td>
<td>3.337</td>
<td>3.451</td>
<td>0.114</td>
</tr>
<tr>
<td>18</td>
<td>1.238</td>
<td>1.258</td>
<td>0.020</td>
<td>0.110</td>
<td>18</td>
<td>2.838</td>
<td>2.831</td>
<td>-0.007</td>
</tr>
<tr>
<td>19</td>
<td>0.889</td>
<td>0.920</td>
<td>0.031</td>
<td>0.083</td>
<td>19</td>
<td>1.493</td>
<td>1.496</td>
<td>0.003</td>
</tr>
<tr>
<td>20</td>
<td>0.598</td>
<td>0.608</td>
<td>0.010</td>
<td>0.074</td>
<td>20</td>
<td>0.891</td>
<td>0.910</td>
<td>0.019</td>
</tr>
</tbody>
</table>

*Note. True = Generating values; Estimates = Mean of point estimates; Bias = Absolute bias; RMSE = Root Mean Squared Error.*

Next, the same quantities for the Part II model are presented in Table 6.3. Recall that we only have two conditions for the Part II model: 20 modified items for Conditions 1 - 16 and 50 modified items for Condition 17 - 20. Hence, we present only the results from Condition 1 for $I = 20$ and Condition 16 for $I = 20$. As in Part I, we only found slight differences within the same conditions in terms of the number of modified items. Previous studies have noted that small cluster sizes and/or large cluster variances tend to result in large error of approximation, especially for binary data (e.g., Cho & Rabe-Hesketh, 2011; Jeon et al., 2017; Joe, 2008). With $I = 20$, there is a clear downward bias
of the variance. When $I$ is increased to 50, the estimated bias and RMSE for both $\mu_c$ and $\sigma_c^2$ improved significantly.

Table 6.3: Estimates of Part II

<table>
<thead>
<tr>
<th>$I$</th>
<th>$\mu_c$ True Estimates Bias RMSE</th>
<th>$\sigma_c^2$ True Estimates Bias RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.287 1.304 0.017 0.278</td>
<td>1.513 1.452 -0.061 0.495</td>
</tr>
<tr>
<td>50</td>
<td>1.296 0.009 0.183</td>
<td>1.503 -0.010 0.289</td>
</tr>
</tbody>
</table>

Note. True = Generating values; Estimates = Mean of point estimates; Bias = Absolute bias; RMSE = Root Mean Squared Error.

Now, we turn to the estimates from Part III, which are of our main interest. For all conditions (Conditions 1 - 20), point estimates, estimated bias, and RMSE of the intercept, regression coefficient, random item effect variance, and random rater effect variance are presented in Table 6.4. Note that the corresponding statistics for the error covariance matrix are omitted to save space given that they yielded both bias and RMSE of essentially 0 across all conditions.

The first observation to be made is that the fixed intercept $\mu$ tended to be downwardly biased, and the bias was generally smaller for smaller $\sigma_\zeta^2$ and $\sigma_\gamma^2$. Specifically, within the same $I$ and $R$, the estimated bias decreased as $\sigma_\zeta^2$ decreased from .25 to .04 given $\sigma_\gamma^2 = .009$ or .04. Similarly, it also decreased as $\sigma_\gamma^2$ decreased from .04 to .009 given $\sigma_\gamma^2 = .04$ or .25. This pattern is not clear for $R = 2$, however. We also found similar pattern for RMSE. Another observation we can make is that increasing the number of raters or increasing the number of items resulted in smaller RMSE. However, increasing the number of raters from 2 to 10 did not help reducing the bias of $\mu$.

Second, the regression coefficient $\pi$ seems to have been well recovered. They were also downwardly biased but not to an appreciable extent. The bias ranged from .004 to .013, and no distinct pattern was found across conditions. RMSE, on the other hand, showed clearly larger values for larger $\sigma_\zeta^2$. Specifically for $I = 20$, RMSE ranged from .040 to .045 when $\sigma_\zeta^2 = .04$, and from .090 to .097 when $\sigma_\zeta^2 = .25$ In addition, similar
to the trend in $\mu$, it appears that RMSE becomes smaller when increasing the number of items (see Conditions 1-4 vs. Condition 17-20), but the pattern is not as clear when increasing the number of raters instead.

Third, the variance of random item effects $\sigma^2_{\xi}$ showed both positive and negative bias. Though not by a large amount, the estimated bias decreased as $\sigma^2_{\xi}$ increased. In contrast, RMSE increased as $\sigma^2_{\xi}$ increased. Note that RMSE weights larger errors more. As expected, the number of raters does not seem to affect estimation, but interestingly, no noticeable pattern was found with respect to the number of items in terms of bias. Similar to $\mu$ and $\pi$, RMSE of $\sigma^2_{\xi}$ reduced when the number of items increased. However, it may not be appropriate to draw a general conclusion given that only four conditions were examined for $I = 50$.

Lastly, the variance of random rater effects $\sigma^2_{\gamma}$ tended to be underestimated. When $\sigma^2_{\gamma}$ was smaller, the bias tended to be smaller as well. The number of raters mattered in the estimation of $\sigma^2_{\gamma}$. When $R = 2$, the largest bias reached -.020, and when the number of raters increased to $R = 10$, it dropped to as low as -.002. RMSE appears to follow the same pattern. As expected, the number of items did not seem to affect the estimation of rater variance at all.
Table 6.4: Estimates of Part III

<table>
<thead>
<tr>
<th>Condition</th>
<th>I</th>
<th>R</th>
<th>σ²_ε</th>
<th>σ²_γ</th>
<th>μ</th>
<th>π</th>
<th>σ²_ε</th>
<th>σ²_γ</th>
<th>Est.</th>
<th>Bias</th>
<th>RMSE</th>
<th>Est.</th>
<th>Bias</th>
<th>RMSE</th>
<th>Est.</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>2</td>
<td>0.04</td>
<td>0.009</td>
<td>0.001</td>
<td>0.001</td>
<td>0.122</td>
<td></td>
<td>0.593</td>
<td>-0.007</td>
<td>0.045</td>
<td>0.046</td>
<td>0.006</td>
<td>0.018</td>
<td>0.005</td>
<td>-0.004</td>
<td>0.010</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
<td>0.002</td>
<td>0.002</td>
<td>0.188</td>
<td></td>
<td>0.592</td>
<td>-0.008</td>
<td>0.043</td>
<td>0.046</td>
<td>0.006</td>
<td>0.018</td>
<td>0.020</td>
<td>-0.020</td>
<td>0.043</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>2</td>
<td>0.25</td>
<td>0.009</td>
<td>0.000</td>
<td>0.000</td>
<td>0.208</td>
<td></td>
<td>0.596</td>
<td>-0.004</td>
<td>0.097</td>
<td>0.250</td>
<td>0.000</td>
<td>0.083</td>
<td>0.008</td>
<td>-0.001</td>
<td>0.034</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>2</td>
<td>0.25</td>
<td>0.04</td>
<td>-0.013</td>
<td>-0.013</td>
<td>0.254</td>
<td></td>
<td>0.598</td>
<td>-0.002</td>
<td>0.095</td>
<td>0.249</td>
<td>-0.001</td>
<td>0.083</td>
<td>0.022</td>
<td>-0.018</td>
<td>0.043</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>3</td>
<td>0.04</td>
<td>0.009</td>
<td>-0.003</td>
<td>-0.003</td>
<td>0.107</td>
<td></td>
<td>0.588</td>
<td>-0.012</td>
<td>0.044</td>
<td>0.044</td>
<td>0.004</td>
<td>0.015</td>
<td>0.007</td>
<td>-0.002</td>
<td>0.008</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>3</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.027</td>
<td>-0.027</td>
<td>0.167</td>
<td></td>
<td>0.587</td>
<td>-0.013</td>
<td>0.045</td>
<td>0.045</td>
<td>0.005</td>
<td>0.016</td>
<td>0.028</td>
<td>-0.012</td>
<td>0.036</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>3</td>
<td>0.25</td>
<td>0.009</td>
<td>-0.015</td>
<td>-0.015</td>
<td>0.198</td>
<td></td>
<td>0.592</td>
<td>-0.008</td>
<td>0.096</td>
<td>0.247</td>
<td>-0.003</td>
<td>0.076</td>
<td>0.008</td>
<td>-0.001</td>
<td>0.009</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>3</td>
<td>0.25</td>
<td>0.04</td>
<td>-0.049</td>
<td>-0.049</td>
<td>0.247</td>
<td></td>
<td>0.592</td>
<td>-0.008</td>
<td>0.096</td>
<td>0.248</td>
<td>-0.002</td>
<td>0.076</td>
<td>0.035</td>
<td>-0.005</td>
<td>0.043</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>4</td>
<td>0.04</td>
<td>0.009</td>
<td>-0.004</td>
<td>-0.004</td>
<td>0.101</td>
<td></td>
<td>0.591</td>
<td>-0.009</td>
<td>0.043</td>
<td>0.045</td>
<td>0.005</td>
<td>0.014</td>
<td>0.007</td>
<td>-0.002</td>
<td>0.007</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>4</td>
<td>0.04</td>
<td>0.04</td>
<td>0.003</td>
<td>0.003</td>
<td>0.141</td>
<td></td>
<td>0.591</td>
<td>-0.009</td>
<td>0.040</td>
<td>0.045</td>
<td>0.005</td>
<td>0.014</td>
<td>0.034</td>
<td>-0.006</td>
<td>0.027</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>4</td>
<td>0.25</td>
<td>0.009</td>
<td>-0.014</td>
<td>-0.014</td>
<td>0.188</td>
<td></td>
<td>0.594</td>
<td>-0.006</td>
<td>0.094</td>
<td>0.249</td>
<td>-0.001</td>
<td>0.076</td>
<td>0.009</td>
<td>0.000</td>
<td>0.008</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>4</td>
<td>0.25</td>
<td>0.04</td>
<td>-0.022</td>
<td>-0.022</td>
<td>0.224</td>
<td></td>
<td>0.594</td>
<td>-0.006</td>
<td>0.093</td>
<td>0.250</td>
<td>0.000</td>
<td>0.077</td>
<td>0.040</td>
<td>0.000</td>
<td>0.035</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>10</td>
<td>0.04</td>
<td>0.009</td>
<td>-0.010</td>
<td>-0.010</td>
<td>0.087</td>
<td></td>
<td>0.592</td>
<td>-0.008</td>
<td>0.040</td>
<td>0.046</td>
<td>0.006</td>
<td>0.015</td>
<td>0.009</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>10</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.023</td>
<td>-0.023</td>
<td>0.111</td>
<td></td>
<td>0.592</td>
<td>-0.008</td>
<td>0.040</td>
<td>0.046</td>
<td>0.006</td>
<td>0.014</td>
<td>0.038</td>
<td>-0.002</td>
<td>0.019</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>10</td>
<td>0.25</td>
<td>0.009</td>
<td>-0.017</td>
<td>-0.017</td>
<td>0.170</td>
<td></td>
<td>0.592</td>
<td>-0.008</td>
<td>0.090</td>
<td>0.251</td>
<td>0.001</td>
<td>0.077</td>
<td>0.010</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>10</td>
<td>0.25</td>
<td>0.04</td>
<td>-0.036</td>
<td>-0.036</td>
<td>0.180</td>
<td></td>
<td>0.594</td>
<td>-0.006</td>
<td>0.093</td>
<td>0.251</td>
<td>0.001</td>
<td>0.077</td>
<td>0.039</td>
<td>-0.001</td>
<td>0.018</td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>2</td>
<td>0.04</td>
<td>0.009</td>
<td>-0.003</td>
<td>-0.003</td>
<td>0.090</td>
<td></td>
<td>0.589</td>
<td>-0.011</td>
<td>0.032</td>
<td>0.044</td>
<td>0.004</td>
<td>0.011</td>
<td>0.005</td>
<td>-0.004</td>
<td>0.009</td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.031</td>
<td>-0.031</td>
<td>0.160</td>
<td></td>
<td>0.589</td>
<td>-0.009</td>
<td>0.031</td>
<td>0.044</td>
<td>0.005</td>
<td>0.011</td>
<td>0.022</td>
<td>-0.003</td>
<td>0.040</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>2</td>
<td>0.25</td>
<td>0.009</td>
<td>0.001</td>
<td>0.001</td>
<td>0.139</td>
<td></td>
<td>0.590</td>
<td>-0.010</td>
<td>0.060</td>
<td>0.246</td>
<td>-0.004</td>
<td>0.053</td>
<td>0.007</td>
<td>-0.002</td>
<td>0.010</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>2</td>
<td>0.25</td>
<td>0.04</td>
<td>-0.044</td>
<td>-0.044</td>
<td>0.231</td>
<td></td>
<td>0.589</td>
<td>-0.011</td>
<td>0.061</td>
<td>0.246</td>
<td>-0.004</td>
<td>0.052</td>
<td>0.038</td>
<td>-0.002</td>
<td>0.068</td>
</tr>
</tbody>
</table>

Note. True = Generating values; Est. = Mean of point estimates; Bias = Absolute bias; RMSE = Root Mean Squared Error
Generating values for mu and coefficient are 0 and 0.6 across all conditions.
6.2.2 Standard Errors

The accuracy of post-convergence standard errors from all replications are reported in Figures 6.1 - 6.5. Specifically, they compare the mean of estimated standard errors against the Monte Carlo standard deviations of the point estimates. As the item parameters in Part I were fixed across replications, those two quantities only slightly vary across conditions are seemingly the same in the plots. Because we only had two distinctive sets of generating parameters for Part II, one for Conditions 1 - 16 and another for Conditions 17 - 20, and so does the corresponding two quantities.

For Part I, we observe quite a close agreement between the mean of the standard errors and the Monte Carlo standard deviations. For Part II, the two values are very close to one another either in Conditions 1 - 16 (see Figure 6.1 - 6.4) or Conditions 17 - 20 (see Figure 6.5), but with a slight downward bias. For Part III, they are also well aligned on the 1:1 line but again with a slight downward bias. The general underestimation of standard errors can be explained by several reasons. First, for efficiency, we only used one imputation per MH-RM cycle in this study. As noted by Cai (2008), such a small number of imputations may lead to estimated standard errors with a downward bias (see multiple imputation theory in Little & Rubin, 1987, for example). Second, we opted for post-convergence standard errors and not recursive standard errors, again for the efficiency reason. While the post-convergence approach improves the stability of estimation for point estimates, the resulting standard errors may be prone to underestimation (e.g., Yang & Cai, 2014), though this may not always be the case in difference research contexts. Lastly, finite sample size could lead to underestimation of standard errors.

6.2.3 Random Effects Scoring: Revised Item Parameters

Our final goal is to obtain revised item parameters. Varying conditions yielded non-identical model parameter recovery. Then the question is whether and to what extent the differential model parameter estimates make a difference in the recovery of revised
item parameters, which corresponds to the random item effects $\mathbf{u}_\alpha$. Pearson correlations, average bias, and average RMSE of the estimated revised item parameters are presented in Table 6.5. Regardless of the conditions examined, $r$ was no less than .99, retaining the rank ordering almost perfectly. In terms of bias, $\mathbf{u}_\alpha$ was biased downwards, as somewhat expected by the underestimation trends observed in the fixed effects parameters (see Table 6.4). Both bias and RMSE tend to decrease with smaller variance of random rater effects $\sigma^2_\gamma$. It was also found that smaller variance of random item effects $\sigma^2_\xi$ reduced RMSE. Surprisingly, increasing the number of raters did not lead to the improvement of final estimates of revised item parameters.

Finally, we examined latent ability recovery with revised item parameters. As an
Figure 6.2: Accuracy of Standard Errors: Conditions 5 - 8

Illustration, the results from Condition 20 is displayed in Figure 6.6. The left panel in Figure 6.6 plots the EAP scores obtained with the estimated revised item parameters ("Estimated") against the generated individual latent ability scores, i.e., $\theta$ ("True"). The individual latent ability scores are well recovered with the estimated revised item parameters with $r = .997$. For all conditions (not displayed here), $r$ was greater than .99, showing good recovery. We still have room for improving the recovery for persons with latent ability at the extreme (above 2 or below 2) by increasing the sample size or including more items that can discriminate those persons.

Additionally, we plotted the EAP scores obtained with the estimated revised item parameters ("Estimated") against the the EAP scores obtained with the "true" revised
item parameters ("Benchmark"), shown on the right panel in Figure 6.6. Unlike item parameters we typically recover in traditional IRT, we newly defined the item parameters in the CCREM that we proposed. One may be interested in how the quality of the estimated item parameters from this model compare to the “true” revised item parameters in terms of latent ability recovery. The EAP scores from the revised item parameter estimates are perfectly aligned with those from the “true” revised item parameters with $r = 1$, all but for a slight underestimation. For all other conditions (not displayed here), $r$ was exactly 1, showing a perfect rank ordering.
Figure 6.4: Accuracy of Standard Errors: Conditions 13 - 16
Figure 6.5: Accuracy of Standard Errors: Conditions 17 - 20
<table>
<thead>
<tr>
<th>Condition</th>
<th>I</th>
<th>R</th>
<th>$\hat{\sigma}_\xi^2$</th>
<th>$\hat{\sigma}_\gamma^2$</th>
<th>$r$</th>
<th>Average Bias</th>
<th>Average RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>2</td>
<td>0.04</td>
<td>0.009</td>
<td>0.993</td>
<td>-0.127</td>
<td>0.182</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
<td>0.994</td>
<td>-0.127</td>
<td>0.227</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>2</td>
<td>0.25</td>
<td>0.009</td>
<td>0.993</td>
<td>-0.123</td>
<td>0.191</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>2</td>
<td>0.25</td>
<td>0.04</td>
<td>0.990</td>
<td>-0.138</td>
<td>0.240</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>3</td>
<td>0.04</td>
<td>0.009</td>
<td>0.996</td>
<td>-0.133</td>
<td>0.176</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>3</td>
<td>0.04</td>
<td>0.04</td>
<td>0.996</td>
<td>-0.155</td>
<td>0.229</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>3</td>
<td>0.25</td>
<td>0.009</td>
<td>0.991</td>
<td>-0.133</td>
<td>0.181</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>3</td>
<td>0.25</td>
<td>0.04</td>
<td>0.993</td>
<td>-0.158</td>
<td>0.233</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>4</td>
<td>0.04</td>
<td>0.009</td>
<td>0.997</td>
<td>-0.132</td>
<td>0.168</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>4</td>
<td>0.04</td>
<td>0.04</td>
<td>0.997</td>
<td>-0.126</td>
<td>0.190</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>4</td>
<td>0.25</td>
<td>0.009</td>
<td>0.996</td>
<td>-0.137</td>
<td>0.174</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>4</td>
<td>0.25</td>
<td>0.04</td>
<td>0.995</td>
<td>-0.147</td>
<td>0.205</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>10</td>
<td>0.04</td>
<td>0.009</td>
<td>0.997</td>
<td>-0.136</td>
<td>0.156</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>10</td>
<td>0.04</td>
<td>0.04</td>
<td>0.998</td>
<td>-0.147</td>
<td>0.175</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>10</td>
<td>0.25</td>
<td>0.009</td>
<td>0.997</td>
<td>-0.140</td>
<td>0.162</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>10</td>
<td>0.25</td>
<td>0.04</td>
<td>0.998</td>
<td>-0.159</td>
<td>0.186</td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>2</td>
<td>0.04</td>
<td>0.009</td>
<td>0.992</td>
<td>-0.137</td>
<td>0.188</td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
<td>0.992</td>
<td>-0.166</td>
<td>0.245</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>2</td>
<td>0.25</td>
<td>0.009</td>
<td>0.991</td>
<td>-0.132</td>
<td>0.193</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>2</td>
<td>0.25</td>
<td>0.04</td>
<td>0.992</td>
<td>-0.177</td>
<td>0.280</td>
</tr>
</tbody>
</table>

*Note.* $r$ indicates Pearson correlation; Average Bias = Average absolute bias across items; Average RMSE = Average Root Mean Squared Error across items.
Figure 6.6: Latent Ability Recovery

Note: "Estimated $\theta$" is the EAP scores obtained with the estimated revised item parameters. "True $\theta$" is the generated individual latent ability scores. "Benchmark $\theta$" is the EAP scores obtained with the "true" revised item parameters.
CHAPTER 7

Discussions

7.1 Summary

This study developed a unified cross-classified random effects model that jointly utilizes item response data from a general population (prior calibrations) and judgmental data from expert raters in order to offer a path to revising the item parameters for use in scoring modified test forms. The present research was motivated by ELP assessment wherein significant barriers exist in producing comparable scores for ELs with disabilities, particularly BLV (blind or low-vision) students, who take different test forms from ELs from the general population. This research follows the recent “judgment-based, data-informed linking process” by Winter et al. (2018) under ELPA21. The method was developed with the aim of applying equivalent cut-scores of the original test forms taken by the general ELs to modified Braille test forms while ensuring the comparability of scores. Since item calibration or the standard linking procedure are not feasible for special populations due to limited sample size, experts’ judgments on the Braille items and prior item calibrations from ELs from the general population were used in conjunction.

More specifically, we introduced a three-part model, a combination of 1) a generalized linear mixed model with fixed item and random person to model responses on the unmodified portion of the original test form; 2) a generalized linear mixed model with cross-classified random effects to model responses on the modified portion of original test form; 3) a multivariate nonlinear cross-classified random effects model with covariates to model judgmental ratings on the Braille test form.
Notably, the joint estimation of the three-part model was performed using the MH-RM algorithm (Cai, 2008, 2010a, 2010b), which has been implemented in R (R Core Team, 2018). After the maximum likelihood (ML) solutions to the cross-classified random effects model were obtained via the MH-RM algorithm, empirical Bayes (EB) prediction was used for random effects scoring to finally estimate the item random effects. This leads to revised item parameter estimates for the modified Braille test forms and ultimately provides comparable scores for BLV students.

The proposed model and its estimation was applied to the data from the Braille ELP assessment in ELPA21, the prime target of this study. Specifically, we focused on the dichotomous item responses for the Listening and Reading domains. The results empirically demonstrate that the standard approach of applying the item parameters from the original version to the Braille version is problematic. That is, we cannot apply the item parameters from the original version to the Braille version. Furthermore, we are left with room for improvement in estimating the revised item parameters within the linear fixed-effects approach. In conclusion, BLV students could have received unfair scale scores had no adjustments in item parameters been made, which can lead to inappropriate and undeserved classifications. It shows that the proposed approach is a promising one for obtaining correct revised item parameters for the BLV forms and ultimately providing comparable scores to those of the general ELs.

The findings were confirmed in preliminary simulation studies via parameter recovery and scoring, though more thorough and extensive simulations are required and additional improvements are desired. Given the limited extent of simulations, central findings can be summarized as the following. First, in general, the parameters were underestimated, which resulted in underestimation of the random item effects (i.e., the revised item parameters). Second, the revised item parameter estimates preserved rank ordering almost perfectly in all conditions, and smaller rater variance reduced the negative bias. Lastly, the IRT scale scores with the estimated revised item parameters were equally well-recovered across conditions, with a slight improvement for larger number
of items. Given the limited study, we can carefully conclude that we may be able to suc-
cessfully obtain revised item parameters with small number of raters as long as raters’
variance is not so large, which is greatly accomplished in the current design of ELPA21.

7.2 Future Directions

As far as we know, this is the first study to propose a method for producing item pa-
rameters from judgmental data for a small-sample linking method with random effects
approach. That being said, three avenues of future research is suggested: extension
of the limited scope both substantively and methodologically, enhancement of the data
collection/study design and modeling, and further scrutiny of the estimation scheme.
These three avenues for future work are detailed in the following.

First of all, extension of scope is needed not only for the generality of this research
itself but also for wide applicability in practice. For one, the simulation conditions need
to be expanded. Herein we focused on an application to the Braille ELP assessment, and
that specifically conducted under ELPA21, and hence the generating conditions largely
relied on this particular context. Examination under a wider range of conditions will
provide more meaningful suggestions for future designs. More broadly, this study can
potentially be extended to any assessment beyond Braille and even beyond ELP assess-
ment. An immediate application would be for alternate assessment for students with
significant cognitive disabilities. Importantly, we limited our discussions to dichoto-
mous items and modified items with source items. While it is straightforward to extend
to polytomous items in theory, it requires additional considerations in modeling and
estimation and on top of further simulation study. New strategies are required for ob-
taining revised item parameters for newly developed items (see Table 3.1) as well, which
do not have counterpart source items.

Second, the proposed model can be extended to develop a better model for moder-
ated item calibration. For example in the current set-up, we assumed a common shift.
That is, we assumed that raters make judgments only by the difficulty parameter. However, any shift in item discrimination parameter from the original assessment can be accommodated further. This is thought to be especially challenging. First, it is of question how successfully raters are able to estimate item discrimination indices. It is very doubtful judging by previous research (Hambleton & Jirka, 2006). Another obstacle stems from statistical aspect. In such a case, additional random effects are required. The interaction the random item discrimination parameter and latent ability is a not trivial to estimate, obviously demanding further research. Another direction we can take is to model other covariates. For instance, we can treat task type as an item covariate and rater’s confidence about his/her ratings as a person covariate to explain the responses or ratings better. This not only helps uncover the relationship between these rather important covariates and the changes in item parameter estimates they would incur, but is also expected to make them more accurate.

Also, one may question whether the design pursued in this work is even the best design. Our discussions have been devoted to the current set-up in ELPA21. There are unanswered questions that we could address in future research. For one, in the current study, we used ratings on the continuous scale. It is questionable however whether even an expert can really distinguish between 71% and 74%, for example. In the data collection stage, we can ask raters to provide categorical responses rather than quantitative statements about difficulty. In addition, we used two cut-scores that are thought to be meaningful in classifying students, which are the same cut-scores as those for the general population determined through a standard setting procedure. What we have not considered is the effect of the number, the locations, or the nature (fixed vs. random) of the cut-scores on item parameter estimation. For example, Cai and Huang (2019) has discussed how the determination of the item parameters may be less justifiable when the items are easy relative to the cut-score locations at which the raters provide their judgements. Their proposed method, which was borrowed from the optimal designs literature (e.g., Sitter & Wu, 1993), may be incorporated into this study’s modeling framework for
investigating the best locations of cut-scores. These factors can be further examined in future research.

Finally, the preliminary experiments of using MH-RM for this unified model via simulation study suggest that it is a promising estimation algorithm for models with cross-classified random effects. Estimation of the Part I model is simply an application of MH-RM on an unidimensional IRT model, which was proposed by Cai (2008, 2010a, 2010b) and has been widely used in practice. However, estimation of the Part II and III models, each corresponding to GLMMs and NLME with crossed random effects, calls for a stand-alone research. Even though MH-RM for GLMMs (including CCREM) has been studied by Chalmers (2015), further exploration focused on CCREM is needed. Different methods of estimation can be compared with Markov chain Monte Carlo (MCMC) as a benchmark. In addition to estimation of this type of model, this study involved difficulties in dealing with the issue of small variances of random effects parameters coupled with small cluster sizes in hierarchical models. A way to overcome or bypass it elegantly would be another independent topic to be studied.


