An Examination of Parameter Recovery Using Different Multiple Matrix Booklet Designs

Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

Dissertation

zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

vorgelegt von
Anta Akuro, M.Ed.

Berlin, 2020

Gutachter/in: 1. Prof. Dr. Martin Brunner

2. Prof Dr. Steffi Pohl

Datum der Einreichung: 13. Januar 2020

Tag der Disputation: 24. April 2020

Acknowledgements

The Acknowledgments is not included in the Online version due to data protection reasons.

Summary

Educational large-scale assessments examine students' achievement in various content domains and thus provide key findings to inform educational research and evidence-based educational policies. To this end, large-scale assessments involve hundreds of items to test students' achievement in various content domains. Administering all these items to single students will over-burden them, reduce participation rates, and consume too much time and resources. Hence multiple matrix sampling is used in which the test items are distributed into various test forms called "booklets"; and each student administered a booklet, containing a subset of items that can sensibly be answered during the allotted test timeframe. However, there are numerous possibilities as to how these booklets can be designed, and this manner of booklet design could influence parameter recovery precision both at global and sub-population levels. One popular booklet design with many desirable characteristics is the *Balanced Incomplete 7-Block* or Youden squares design. Extensions of this booklet design are used in many large-scale assessments like TIMSS and PISA. This doctoral project examines the degree to which item and population parameters are recovered in real and simulated data in relation to matrix sparseness, when using various *balanced incomplete block* booklet designs. To this end, key factors (e.g., number of items, number of persons, number of items per person, and the match between the distributions of item and person parameters) are experimentally manipulated to learn how these factors affect the precision with which these designs recover true population parameters. In doing so, the project expands the empirical knowledge base on the statistical properties of booklet designs, which in turn could help improve the design of future large-scale studies.

Generally, the results show that for a typical large-scale assessment (with a sample size of at least 3,000 students and more than 100 test items), population and item parameters are

recovered accurately and without bias in the various multi-matrix booklet designs. This is true both at the global population level and at the subgroup or sub-population levels. Further, for such a large-scale assessment, the match between the distribution of person abilities and the distribution of item difficulties is found to have an insignificant effect on the precision with which person and item parameters are recovered, when using these multi-matrix booklet designs.

These results give further support to the use of multi-matrix booklet designs as a reliable test abridgment technique in large-scale assessments, and for accurate measurement of performance gaps between policy relevant subgroups within populations. However, item-position effects were not fully considered, and different results are possible if similar studies are performed (a) with conditions involving items that poorly measure student abilities (e.g., with students having skewed ability distributions); or, (b) simulating conditions where there is a lot of missing data because of non-response, instead of just missing by design. This should be further investigated in future studies.

Zusammenfassung

Die Erfassung des Leistungsstands von Schülerinnen und Schülern in verschiedenen Domänen durch groß angelegte Schulleistungsstudien (sog. *Large-Scale Assessments*) liefert wichtige Erkenntnisse für die Bildungsforschung und die evidenzbasierte Bildungspolitik. Jedoch erfordert die Leistungstestung in vielen Themenbereichen auch immer den Einsatz hunderter Items. Würden alle Testaufgaben jeder einzelnen Schülerin bzw. jedem einzelnen Schüler vorgelegt werden, würde dies eine zu große Belastung für die Schülerinnen und Schüler darstellen und folglich wären diese auch weniger motiviert, alle Aufgaben zu bearbeiten. Zudem wäre der Einsatz aller Aufgaben in der gesamten Stichprobe sehr zeit- und ressourcenintensiv. Aus diesen Gründen wird in Large-Scale Assessments oft auf ein *Multi-Matrix Design* zurückgegriffen bei dem verschiedene, den Testpersonen zufällig zugeordnete, Testheftversionen (sog. *Booklets*) zum Einsatz kommen. Diese enthalten nicht alle Aufgaben, sondern lediglich eine Teilmenge des Aufgabenpools, wobei nur ein Teil der Items zwischen den verschiedenen Booklets überlappt. Somit wird sichergestellt, dass die Schülerinnen und Schüler alle ihnen vorgelegten Items in der vorgegebenen Testzeit bearbeiten können. Jedoch gibt es zahlreiche Varianten wie diese Booklets zusammengestellt werden können. Das jeweilige Booklet Design hat wiederum Auswirkungen auf die Genauigkeit der Parameterschätzung auf Populations- und Teilpopulationsebene. Ein bewährtes Booklet Design ist das *Balanced-Incomplete-7-Block Design*, auch *Youden-Squares Design* genannt, das in unterschiedlicher Form in vielen Large-Scale Assessments, wie z.B. TIMSS und PISA, Anwendung findet. Die vorliegende Arbeit untersucht sowohl auf Basis realer als auch simulierter Daten die Genauigkeit mit der Item- und Personenparameter unter Anwendung verschiedener Balanced-Incomplete-Block Designs und in Abhängigkeit vom Anteil designbedingt fehlender Werte geschätzt werden können. Dafür wurden verschiede Designparameter variiert (z.B. Itemanzahl, Stichprobenumfang, Itemanzahl pro Booklet,

Ausmaß der Passung von Item- und Personenparametern) und anschließend analysiert, in welcher Weise diese die Genauigkeit der Schätzung von Populationsparametern beeinflussen. Die vorliegende Arbeit hat somit zum Ziel, das empirische Wissen um die statistischen Eigenschaften von Booklet Designs zu erweitern, wodurch ein Beitrag zur Verbesserung zukünftiger Large-Scale Assessments geleistet wird.

Die Ergebnisse der vorliegenden Arbeit zeigten, dass für ein typisches Large-Scale Assessment (mit einer Stichprobengröße von mindestens 3000 Schülerinnen und Schülern und mindestens 100 Items) die Personen- und Itemparameter sowohl auf Populations- als auch auf Teilpopulationsebene mit allen eingesetzten Varianten des Balanced-Incomplete-Block Designs präzise geschätzt wurden. Außerdem konnte gezeigt werden, dass für Stichproben mit mindestens 3000 Schülerinnen und Schülern die Passung zwischen der Leistungsverteilung und der Verteilung der Aufgabenschwierigkeit keinen bedeutsamen Einfluss auf die Genauigkeit hatte, mit der verschiedene Booklet Designs Personen- und Itemparameter schätzten.

Die Ergebnisse untermauern, dass unter Verwendung von multi-matrix Designs bildungspolitisch relevante Leistungsunterschiede zwischen Gruppen von Schülerinnen und Schülern in der Population reliabel und präzise geschätzt werden können. Eine Einschränkung der vorliegenden Studie liegt darin, dass Itempositionseffekte nicht umfassend berücksichtigt wurden. So kann nicht ausgeschlossen werden, dass die Ergebnisse abweichen würden, wenn (a) Items verwendet werden würden, welche die Leistung der Schülerinnen und Schüler schlecht schätzen (z.B. bei einer schiefen Verteilungen der Leistungswerte) oder (b) hohe Anteile an fehlenden Werten vorliegen, die nicht durch das Multi-Matrix Design erzeugt wurden. Dies sollte in zukünftigen Studien untersucht werden.

Contents

Acknowledgements iii
Summary v
Zusammenfassung vii
Contents ix
List of Figures xii
List of Tables xiii
Nomenclature xiv
Chapter 1 Introduction 1
1.1 Large-scale assessments in Education 1
1.2 Multiple matrix booklet designs 5
1.3 IRT and parameter recovery accuracy 8
1.4 Aims and scope 12
Chapter 2 The Rasch Item response Theory Model 15
2.1 Introduction to item response theory 15
2.1.1 The dichotomous Rasch Model 16
2.1.2 Mixed coefficient multinomial logit model (MCMLM) 19
Chapter 3 Estimation of person and item parameters 22
3.1 Maximum likelihood method 22
3.2 Plausible values (PVs) imputation approach 26
3.3 Efficiency measurement based on item response theory 28
Chapter 4 Missing data in planned sampling plans 31
4.1 Planned sampling plans 31
4.1.1 Item Sampling 32
4.1.2 Item-Examinee Sampling 33
4.2 Missing data theory in sampling plans 36
4.2.1 Missing data mechanisms 36

4.2.2 Missing data treatments 40
Chapter 5 Study I—Effect of test length, sample size and population subgroups on measurement. 44
5.1 Background 44
5.1.1 Test length and measurement precision or efficiency 44
5.1.2 Sample size and measurement efficiency 46
5.1.3 Measurement efficiency in policy relevant population subgroups 48
5.2 Research objectives and research questions 51
5.3 Data and procedure 52
5.4 Method of data analyses 58
5.5 Results and discussion 62
5.5.1 Item and person parameter recovery at the global population level 62
5.5.2 Test length and sample size and parameter recovery efficiency 71
5.5.3 Parameter recovery efficiency in policy relevant population subgroups 78
Chapter 6 Study II—Item-person match and parameter recovery efficiency 85
6.1 Item and test information functions 85
6.2 Empirical studies on item person-match 88
6.3 Research objectives and research questions 89
6.4 Data and procedure 90
6.5 Method of data analyses 92
6.6 Results and discussion 93
6.6.1 Item-person match and efficiency of mean person ability estimate recovery 93
6.6.2 Item-person match and efficiency of variance (of person abilities) recovery 100
6.6.3 Item-person match and recovery of mean item difficulty 106
Chapter 7 General discussion 116
7.1 Summary of studies 116
7.2 Discussion of findings 118
7.2.1 Test length, sample size, and parameter recovery efficiency in sparse matrix booklet designs 119
7.2.2 Group level parameter recovery 125
7.2.3 Item-person match and parameter recovery 127
7.3 Implications of research findings 130
7.3.1 Implications for test developers and measurement experts 130
7.3.2 Implications for policy makers, politicians and other stakeholders in Education 132

7.4 Study limitations and recommendations for future research 134
7.5 General conclusion 137
References 139
Appendix A Additional Results 162
A.1 Recovery of person ability distributions across various booklet designs for the set of first six plausible values 163
Appendix B Program Code 164
Erklärung 193
Curriculum vitae 194

List of Figures

Figure 1.1– Different booklet designs 6
Figure 1.2– Difference between low bias and high bias during parameter recovery 10
Figure 2.1– An example of an item characteristic curve 17
Figure 4.1– A fractional block design for incomplete measurements 35
Figure 5.1– Overview of the various booklet designs used in the study 58
Figure 5.2– Recovery of person ability distributions 65
Figure 5.3– Recovery of mean θ and var θ in real data 66
Figure 5.4– Recovery of mean δ across booklet designs 67
Figure 5.5– RMSEs for recovered item locations at item level in VERA.8 dataset 69
Figure 5.6– Bias for recovered item locations at item level in VERA-8 dataset 70
Figure 5.7– RMSE of recovered person and item parameters across booklet designs 73
Figure 5.8– Bias of recovered subgroup mean difference in person ability 74
Figure 5.9– RMSE and bias of recovered mean group difference in θ between subgroups 81
Figure 5.10– RMSE and bias of recovered difference in variance of θ beween subgroups 83
Figure 6.2– RMSE for the recovery of mean θ across all experimental conditions 97
Figure 6.3– Distribution of recovered θfor different item-person match conditions 98
Figure 6.4– RMSE of recovered variance in θ for different item-person match conditions 104
Figure 6.5– RMSE for recovery of mean β across all experimental conditions 111
Figure 6.6– Residuals between true and estimated δ's for 1000 simulations 112
Figure 6.7– RMSE for recovery of δ for individual items across 1000 simulaitons 113
Figure 6.8– Bias for the recovery of δ across 1000 simulations 114
Figure 7.1– Distribution of recovered varaince of θ for two conditions 124
Figure A.1– Recovery of θ ability from first six plausible values 163
Figure A.2– Bias for recovery of mean θ across person-item match conditions 164
Figure A.3– RMSE for recovery of mean θ across person-item match conditions 165
Figure A.4– Bias for recovery of var θ across person-item match conditions 166
Figure A.5– Bias for recovery of mean δ across person-item match conditions 167

List of Tables

Table 5.1 – Overview of the study design 54
Table 5.2 – Fit statistics for VERA-8 items 56
Table 5.3 – Parameter recovery efficiency for item parameters across booklet designs 64
Table 5.4 – RMSE and bias for recovery of δ in VERA-8 dataset 68
Table 5.5 – RMSE of recovered θ and δ across test lengths and sample sizes 75
Table 5.6 – Bias for recovered θ and δ across test lengths and sample sizes 76
Table 5.7– Summary of ANOVA for recovery of person and item parameters 77
Table 5.8 – RMSE and bias of recovered group difference in θ for population subgroups 82
Table 5.9 – RMSE and bias recovered group difference in θ for population subgroups 84
Table 6.1 – Match conditions between distributions θ and δ in simulated data 91
Table 6.2 – Overview of the study design 92
Table 6.4 – RMSE of recovered θ across various all conditions investigated 95
Table 6.5 – Bias of recovered θ across all conditions investigated 96
Table 6.6 – Summary of ANOVA with log(RMSE) of recovered mean θ as criterion 99
Table 6.7 – RMSE of recovered variance of θ across all experimental conditions 102
Table 6.8 – Bias of recovered mean δ across all experimental conditions 103
Table 6.9 – Summary of ANOVA with log(RMSE) of recovered variance θ as criterion 105
Table 6.10 – RMSE of recovered mean δ across experimental conditions 109
Table 6.11 – Bias of recovered mean δ across experimental conditions 110
Table 6.12 – Summary of ANOVA with log(RMSE) of recovered mean δ as criterion 115
Table 7.1 – Sample size recommendations from some IRT studies on parameter recovery 120

Nomenclature

1-PL	1-Prameter Logistic
2-PL	2-Parameter Logistic
3-PL	3-Parameter Logistic
BIBD	Balanced Incomplete Block Design
CAT	Computer Adaptive Testing
CML	Conditional Maximum Likelihood
CMS	Composite Measurement Scale
CONFEMEN	Conference of the Ministers of Education of French Speaking Countries
EAP	Expected A Posteriori
ETS	Education Testing Service
GCSE	General Certificate of Secondary Education
GPA	Grade Point Average
ICC	Item Characteristic curve
IRT	Item Response Theory
JML	Joint Marginal Likelihood
LSA	Large Scale Assessment
MCAR	Missing completely At Random
MAR	Missing at Random
MCMLM	Mixed Coefficient Multinomial Logit Model
MIRT	Multidimensional Item Response Theory
OECD	Organization for Economic Cooperation and Development
PACEC	Programme for the Analysis of Education Systems
PCM	Partial Credit Model
PILNA	Pacific Islands Literacy and Numeracy Assessment
PISA	Programme for International Student Assessment
RMSE	Root Mean Square Error
SACMEQ	Southern and Eastern African Consortium for Monitoring Education Quality
SAT	Scholastic Achievement Test
SQS	Survey Questionnaire Sampling
TAM	Test Analysis ,Modules
TIMSS	Trends in International Mathematics and Science Study
TMST	Targeted multistage testing
UNESCO	United Nations Educational, Scientific and Cultural Organization
VERA 8	“Vergleichsarbeiten in der 8. Jahrgangsstufe”

Indices

θ	Person ability
δ	Item difficulty

Chapter 1 Introduction

Large-scale assessments provide key findings to inform educational research and evidence-based educational policies. Multiple matrix booklet designs in conjunction with item response theory models form a bedrock to the state-of-the-art methodology in these assessments. However, a central issue with data treated with multi-matrix designs and item response models, is precision of estimated parameters. Factors such as test length, sample size, matrix sparseness in booklet design, and item-person match could affect the precision with which item and population parameters are recovered when using these multiple matrix booklet designs. It thus becomes important to investigate conditions under which very accurate item and population parameters are recovered (both at the global and subpopulation levels), when using these multi-matrix booklet designs.

This chapter begins with a discussion of what large-scale assessments are. This will involve, a description of several large-scale educational assessments—applied both at national and international levels. This will be followed by a brief summary on multi-matrix booklet designs as applied in large-scale assessments. Particularly, emphasis will be given to the balanced incomplete block design, which is a popular multi-matrix booklet design, and used in many large-scale assessments like PISA. The chapter will thus end with a brief discussion on the issue of parameter recovery accuracy in item response models; and, a summary of the aims and scope of this doctoral project.

1.1 Large-scale assessments in Education

According to data collected by UNESCO in 2006 and 2007, large-scale educational assessments are becoming a rapidly growing phenomenon in virtually all world regions, with the number of countries carrying out these assessments more than doubling from 28 to 67

between the years 1995 and 2005 (Benavot & Tanner, 2007). Further, although developed countries continued having the highest participation rates, developing countries almost doubled their rate of participation from 28 to 51 percent (Benavot & Tanner, 2007).

In the broadest sense, large-scale assessments can be considered as surveys of knowledge, skills, or behaviours in a given domain, with an objective to describe a population(s) of interest, for instance countries, states or regions (Kirsch et al., 2013; Cordero, Christobal & Santin, 2017). Simon, Ercikan, and Rosseau (2013) define them as standardized assessments conducted on a regional, national or international scale and involving large populations. The assessments focus on group scores and can be differentiated from large-scale testing programs (like the General Certificate of Secondary Education, GCSE; or the Scholastic Achievement Test, SAT) which focus on assessing individuals.

Initially, their function was to help examine students' grades in their academic courses, and to act as a monitor of provincial education systems (Klinger, DeLuca, & Miller, 2008). More recently, they have become more widespread—with many provinces in the United States using them for educational system accountability (Klinger & Rogers, 2011; Linn, 2003). One explanation for such growing interest in large-scale assessments at the provincial and state level is the need for policymakers to find tools for gathering information about their own system's performance—considering increased globalization—and, the common belief that these assessments are necessary to bring about change to improve the quality of schools and student learning (Dolin, 2011).

A very classic example of a large-scale assessment at the national level is the National Assessment of Educational Progress (NAEP) carried out in the United States. Since the late 1960's, samples of students within the U.S. have taken part in this assessment. NAEP is administered to children at the fourth, eighth and twelfth grades, and covers a wide range of subjects like mathematics, science, reading, writing, history and civics (Naemi et al., 2013). Other examples of large-scale assessments at the national level are the evaluation of national standards, i.e., the "IQB-Ländervergleich" and the "IQB-Bildungstrend" in Germany (Pant, Stanat, Schroeders, Roppelt, Siegel, & Pöhlmann, 2013); and the Pan-Canadian Assessment Program (PCAP) in Canada (Gonzalez & Rutkowski, 2010).

At the international level, the most popular large-scale assessment studies (Koehler, 2015) are the Programme for International Student Assessment (PISA), the Third International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). PISA assesses the mathematics, science and reading performance of 15-year-olds every three years since the year 2000. TIMSS assesses the mathematics and science achievement of fourth and eighth graders every four years since 1995; while PIRLS focuses on the reading literacy of fourth graders, who have been surveyed every five years since 2001 (Cordero et al, 2017). Other less popular international large-scale assessments include (Tobin, Lietz, Nugroho, Vivekanandan, & Nyamhuu, 2015): The Southern and Eastern African Consortium for Monitoring Educational Quality (SACMEQ); Conference of the Ministers of Education of French Speaking Countries' (CONFEMEN), Programme for the Analysis of Education Systems (PASEC); Latin American Laboratory for the Assessment of the Quality of Education; and, Pacific Islands Literacy and Numeracy Assessment (PILNA).

SAQMEQ was created in 1995 and consists of a voluntary and collaborative grouping of 15 ministers of education from southern and eastern African states[1]. The education ministry of each participating country collects information on baseline indicators for educational inputs, general conditions of schooling, equity assessments for human inputs, material resource allocations, and literacy levels among grade 6 students (Kellaghan & Greaney, 2008). PASEC on the other hand was established in 1991 as a network for sharing educational evaluation instruments and results amongst French speaking African countries[2]. Initially, this assessment collected data for pupils in grades 2 and 5 only; though, this has been expanded to include pupils in all grades from grades 2 to 6, as well as additional background data on teachers and a variety of school factors (Kulpoo & Coustère 1999). Similarly, PILNA has been conducted twice (in 2012 and 2015) by 13 pacific countries; with the aim of establishing a regional baseline for the literacy and numeracy achievement of students at the end of Year 4 and Year 6 (Belisle, Cassity, Kacilala, Seniloli, & Taoi, 2016).

[1] These countries include Botswana, Lesotho, Kenya, Malawi, Mauritius, Seychelles, South Africa, Swaziland, Tanzania, Tanzania-Zanzibar, Uganda, Zambia, and Zimbabwe (Heyneman & Lee, 2014)
[2] These francophone countries include Mauritania, Cameroon, Senegal, Cape Verde, Guinea Bissau, Guinea, the Ivory Coast, Togo, Benin, Burkina-Faso, Niger, Central African Republic, Congo (Brazzaville), Gabon, Madagascar, Comoros, the Seychelles, Mauritius, Djibouti, and Burundi (Heyneman & Lee, 2014)

One interesting development in large-scale assessments is that they no longer collect information on cognitive measures only; but have expanded in scope to collect information on non-cognitive outcomes and skills (Kaplan & Su, 2016). This is achieved using context or background questionnaires which provide important information on variables used in models to predict cognitive outcomes. The reason for this development could be increasing concerns about the distribution of human capital, and growing recognition that such non-cognitive skills also contribute to the prosperity of individuals and nations (Kirsch, Lennon, von Davier, Gonzalez & Yamamoto, 2013).

Further, the rise in participation rates at LSA's has resulted in a drift from their traditional use of examining differences between educational systems—to evaluating how education services are delivered and the outcomes from such delivery (Kamens & McNeely, 2009). They keep track of the education outcomes for examinees in particular sub-groups, especially those that have been known to suffer educational disadvantages—like boys (Younger & Warrington, 2005; Hannover & Kessels, 2011); children from poor socio-economic backgrounds (APA Task Force on Socioeconomic Status, 2007; Bradley & Corwyn, 2002); or, children from rural or less-developed areas (Roscigno & Crowle, 2009)—and use this to inform initiatives aimed at addressing such inequity. Large-scale assessments therefore offer a means of giving a common reference to all stakeholders involved with an education system. A common benchmark is used in the assessment ensuring that every student is measured in the same way. This is unlike the case of classroom assessments where students in different schools are sometimes tested with different instruments hence disfavouring fairness and equity.

Large-scale assessments are therefore indispensable for any data-driven or student-centred education system since, they provide data that increase policy-makers' understanding of crucial school and non-school factors that may impact teaching and learning; serve as resource for finding areas of concern and action used in preparing and evaluating resulting educational reforms; as well as, play a key role in developing and improving the capacity of education systems to partake in national programmes for educational monitoring and improvement (Rutkowski & Gonzalez, 2010; Wagemaker, 2014).

1.2 Multiple matrix booklet designs

Multiple matrix booklet designs in combination with IRT (Item Response Theory) analyses represent the state-of-the-art methodology used in large-scale assessment studies. Typically, in these large-scale assessments, several hundred questions are used to measure students' performance or achievement in several content domain(s). This ensures sufficient construct representation (of broad skill domains, like mathematics, reading or science); which in turn leads to better content validity since elements of the assessment instrument become more relevant and representative of the targeted construct for the assessment (Haynes, Richard & Kubany, 1995). For instance, in PISA, about 150 to 200 items are used to measure students' achievement in mathematics, reading, and science in each assessment (Frey & Bernhardt, 2012). With such a large number of test items in a LSA, presenting every question to each test taker could over burden them, reduce participation rates, increase administration costs, or even take too much time (Wolf 2006).

To remedy this situation, large scale assessments utilize multiple matrix sampling, wherein every examinee is presented only a subset of overall test items. Before administering these items, they are distributed into test forms (known as "booklets" in large scale assessment terminology), with every booklet containing only an adequate number of items an examine can sensibly answer within the allotted test duration. The manner of distributing these items into booklets is referred to as a multiple matrix booklet design (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; Yousfi & Böhme, 2012). Also, having constructed all booklets, examinees are presented with one booklet; and, though each examinee answers only a subset of the entire test, after collecting all booklets from all examinees, it becomes possible to get information on all items in the overall test.

Multi-matrix sampling (Shoemaker, 1973), or, item-sampling (Lord, 1962) in older literature, is therefore the sampling technique used generally in booklet designs and comes from the procedure of sampling both items and examinees; that is, giving a subset of items to a subset of examinees (Gonzalez & Ruthowski, 2010). In other words, it is a method of assembling and administering a survey or assessment where each respondent is measured on a sample of

the total assessment (Rutkowski, Gonzalez, von Davier, & Zhou, 2014). For example, a Mathematics proficiency test consisting of 40 items could be sub-divided into four subsets of ten items each; and samples of the population of students are each randomly given three of the item subsets to answer. This implies that each student will answer thirty out of the total of 40 test items. Moreover, although each examinee tested is presented with only a portion of the total number of 40 items, the results from each subtest may be used to estimate the statistic that could have been obtained—from the complete test—that is, in the case where all 40 items are given to all participating students. (Gressard & Loyd, 1991, Shoemaker, 1973).

Although multiple matrix booklet designs are constructed to match each testing situation, Gonzalez and Rutkowski (2010) describe two major classifications: complete and incomplete multi-matrix booklet designs. In the complete designs (Figure 1.1a) each booklet contains all different blocks of the test (A, B, and C), meaning each student answers all items in the complete test. The advantage is that by rotating the various blocks across booklets, block order effects could be checked, and examinees prevented from copying from one another.

Figure 1.1. Different Booklet Designs: (a) Complete (b) incomplete booklet design with each comprising three booklets and (c) Balanced incomplete block (Youden squares) design.

(a) Complete booklet design

Booklet	Item Blocks		
1	A	B	C
2	B	C	A
3	C	A	B

(b) Incomplete booklet design

Booklet	Item Blocks	
1	A	B
2	B	C
3	C	A

(c) Balanced incomplete block (Youden squares) Design

Booklet	Item Blocks		
1	A	B	D
2	B	C	E
3	C	D	F
4	D	R	G
5	E	F	A
6	F	G	B
7	G	A	C

Note. Figure adapted from "Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments" by E. Gonzalez and L. Rutkowski (2010*), IERI Monograph Series: Issues and Methodologies in large-scale assessments*, 3, p. 136-137. Copyright 2010 by Educational Testing Service (ETS).

The disadvantage is however that since each student answers all assessment items, the design offers no reduction in respondent burden (i.e., total number of items an examinee ought to answer). Because of this shortcoming, large-scale assessments seldom make use of *complete* multi-matrix booklet designs, where minimizing respondent burden is often a much-desired objective.

On the other hand, incomplete booklet designs (Figure 1.1b) are constructed such that each *booklet* does not contain all the blocks in the test. In the incomplete block in Figure 1.1b, each booklet contains only two blocks allowing for test time to be reduced by 33 percent. A major disadvantage of the design being that correlations cannot be calculated between all pairs of item blocks. This short coming is removed by the *balanced incomplete 7-block* (BIB-7) or Youden squares design (Preece, 1990; Johnson, 1992). BIB-7 designs can only be constructed with item blocks that are a multiple of 7, thus the use of seven item blocks (i.e., item blocks A to G) in the BIB-7 design in Figure 1.1c above. The balance ensures that each block of items appears an equal number of times in each position across the entire booklet design. This makes controlling for two factors possible (booklet and item or cluster position—in large scale-assessment terminology) which could have an undesirable impact on relevant parameter estimates (Frey, Hartig & Rupp, 2009; Frey & Bernhardt, 2012).

Further, since each item pair occurs together an equal number of times, it makes it feasible to get a full inter-item correlation matrix necessary for computing parameter estimates; and improves measurement precision, since design balance and replication reduce standard deviation and variability of recovered item and person parameters (Cochran & Cox, 1992). Because of these numerous advantages, the BIB design was successfully implemented for the first time in an educational setting at the 1983/1984 NAEP assessment (Beaton, 1987; Beaton & Zwick, 1992; Johnson, 1992); with extensions of the design used in important large-scale assessments like TIMSS and PISA (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; Rutkowski, Gonzalez, von Davier, & Zhou, 2014).

An extended balanced incomplete block design according to Frey, Hartig and Rupp (2009) is one in which: (1) Every cluster of items (t) occurs at most once in a booklet (b). (2) Every cluster appears equally often (r) across all booklets. (3) Every booklet is of identical length,

containing the same number of clusters (k). (4) Every pair of clusters occurs together in the booklets with equal frequency (λ). Thus, t, b, r, k, and λ are called the parameters of the design and characterize any extended incomplete block design. For example, the *balanced incomplete 7-block* as displayed in Figure 1c above is characterized by the parameters t = 7 clusters, b = 7 booklets, r = 3 occurrences in each booklet, and $\lambda = 1$ occurrence of each cluster pair.

Multiple matrix designs (especially the balanced incomplete block designs) therefore play a vital role in the current methodology of large-scale assessments, since they help in the leverage of resources by permitting fewer items to be answered per student, while allowing yet so many questions to be asked in the test to cover broad content domains. Such a technique of test construction and design is important since—in principle—it allows the estimation of achievement distributions for target populations and sub-populations; and full coverage of the assessment framework, while also simultaneously keeping examinee burden and testing time at the school reduced (Gonzalez & Rutkowski, 2010). Their disadvantage however is, they are unsuitable when estimating individual student proficiencies since each student answers not enough items to ensure sufficient test score reliability (Rutkowski, Gonzalez, von Davier, & Zhou, 2014).

1.3 IRT and parameter recovery accuracy

Parameter recovery refers to how well an estimate of a population parameter is obtained. This parameter could be a person parameter (for instance, a person's ability on a latent trait), or an item parameter (such as an item's difficulty). Further, Item response theory (IRT) serves as the modern statistical framework for handling fundamental testing challenges imposed by multi-matrix designs. Other important test settings where this framework is applied include determining examinee proficiency for certification purposes; test item assembly; equating different tests; and, examining potential bias that test items could express towards certain minority or focal groups (Swaninathan, Hambleton, Sireci, Xing, & Rivazi, 2003). IRT models make it possible to describe the probability of giving a correct response to an item based on the underlying ability of an examinee (i.e., the person parameter) and item difficulty

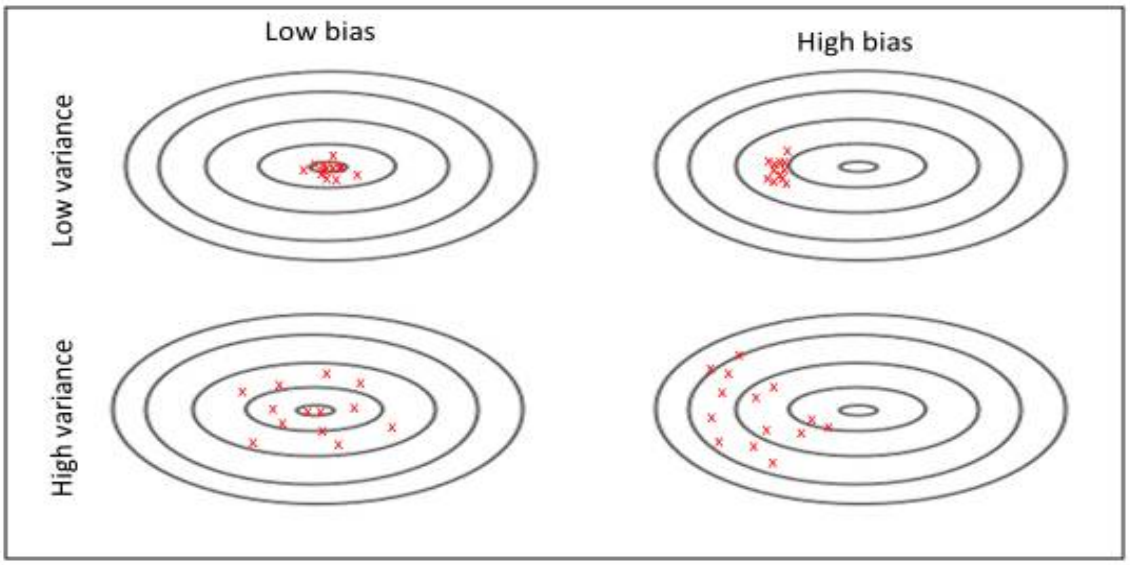
(i.e., the item parameter) (Foley, 2010). The main benefit of this measurement framework is that, (when the model fits reasonably well to the data) estimates of item parameters are examinee or sample independent; and, estimates of person ability independent of the items (Hambleton, Swaninathan, & Rogers, 1991). This is not the case with test-based classical test theory, where there is no possibility to predict how an examinee will perform on a given test item. Further, classical test theory (CTT) also known as the true score theory, is based on the idea that each person has a true score, *T*, which would be obtained in the absence of any error in measurement (Cappelleri, Lundy, & Hays, 2014). IRT is therefore often preferred over CTT because it provides greater flexibility—as a broader range of interpretations are made at the item level; and, permits to predict the likelihood of a given examinee answering any chosen item correctly (Hambleton & Jones, 1993).

However, to get the full advantages IRT offers, it is of utmost importance to ensure that IRT model parameters—person and item parameters—be accurately estimated. As emphasized by Kieftenbeld and Natesan (2012, p.399), "Accurate recovery of model parameters from response data is a central problem in item response theory." An important requirement to utilizing IRT models is therefore ensuring the accuracy and stability of model parameters (He & Wheadon, 2013). Stability and accuracy play a key role in the development and design of IRT-based tests.

Accuracy and stability of parameter recovery is typically examined using bias and root mean squared error (RMSE) between the estimated and true parameters. Bias and RMSE are chosen because these are the most popular indices used in studies examining parameter recovery accuracy with item response models (e.g., see Custer, 2015; Svetina et al., 2013; Hecht, Weirich, Siegle, & Frey, 2015a; Toland, 2008). The bias describes the mean difference of the estimated parameters and the true parameters—in other words, mean inaccuracy of the parameter estimate. The RMSE is the root of the averaged squared difference between the estimated and corresponding true parameter. Hence, the RMSE takes the variability of the estimate into account—with smaller values for bias and smaller values of the RMSE indicating better parameter recovery. The lower the variance of estimated values of a given parameter, the lower the resulting RMSE irrespective of the direction of the

variance; however, a high variance could result in low bias, if the recovered parameter estimates lie on both sides of the true parameter value—thus cancelling out (See Figure 1.2).

Figure 1.2. Difference between low bias and high bias during parameter recovery



Note. Predicted or estimated values may differ from the true scores in two ways, (a) being biased by systematically deviating from the true scores, or (b) portraying an unsystematic but high degree of imprecision or variance. Figure adapted from "Choosing prediction over explanation in Psychology: Lessons from machine learning" by T. Yarkoni and J. Westfall, 2017, *Perspect Psychol Sci*, 12(6), p. 1105. Copyright 2017 by The Authors.

Importantly, accuracy and stability of parameter recovery in IRT modelling could be affected by a myriad of conditions and factors like researcher's choice among IRT models, estimation methods, software programs, test length, sub-groups within the population, and shapes of item and person parameter distributions (Hambleton, 1989; Wollack, Sung, & Kang, 2006). Further, most studies investigating the influence of these factors on parameter recovery when using item response models use simulated data (e.g., Svetina, Crawford, Levy, Green, Scott, Thompson, Gorin, Fay, & Kunze, 2013; Jiang, Wang, & Weiss, 2016; Montgomery & Skorupski, 2012). Importantly, extremely few studies have investigated this subject when using multi-matrix designs[3].

[3] This was validated using a google search done in March 2019 with the keywords: (1) multi-matrix designs, parameter recovery accuracy; (2) multiple matrix booklet designs, parameter recovery precision

Using multi-matrix designs is important because, they involve sparse or missing data and effects that were negligible when using complete data could become significant when using sparse data. For example, in one of the recent studies examining parameter recovery when using an IRT model, Svetina et al. (2013) carried out a simulation to investigate how the match between person and item parameter distributions influenced parameter recovery when short tests were given to small samples. The data used was complete (with no missingness); and factors manipulated were the match between person and item parameter distributions, test length, sample size, and item discrimination. Their results showed that mismatch between person and item parameter distributions had little impact on parameter recovery; and that parameter estimation accuracy reduced as sample size and test length became smaller.

Moreover, the question on parameter recovery and match between item difficulty and student ability distributions is interesting because, most large-scale assessments assume both distributions to be standard normal (i.e., with a mean of 0, and a standard deviation of 1) whereas, it is possible to have populations where this is not true—for instance, one region or country in a large-scale assessment having students with very high abilities, with the mean ability being largely greater than 1; or having students with very low ability, with a mean ability greatly less than one. Such a situation clearly results in a mismatch in the ability distribution and item difficulty distribution of students from such regions or countries; and could have undesirable effects on recovery of person or item parameters.

Further, though multi-matrix designs play such a key role in large-scale assessments, the empirical knowledge base on multi-matrix booklet designs and parameter recovery is still very limited, with “much of the discussions around multi-matrix sampling having been relegated to the pages of technical manuals” (Rutkowski, Gonzalez, von Davier, & Zhou, 2014, p.76). This limited knowledgebase includes the studies of Gressard and Loyd (1991) and that of Gonzalez & Rutkowski (2010).

Gressard and Loyd (1991) used achievement data and a Monte Carlo approach to investigate the effect of item sampling by item stratification on parameter estimation, when using different multi-matrix booklet designs. The designs were created based on matrix sparseness (i.e., total number of items answered per student), with each design having a different number

of subtests and items per subtest. Their results showed that the item sampling method and booklet design which is a practical compromise in terms of precision and sample size is one where the items are stratified with respect to how well they distinguished between high and low achieving students; and the sampling plan that has a modest number of subsets. This sampling condition gives reasonable precision of the mean and variance estimates but needs only a moderately sized sample.

Gonzalez and Ruthowski (2010) used balanced incomplete block designs[4] to carry out a simulation study. They examined the extent to which item and population parameters were recovered given sample size and matrix sparseness; and simulated mathematics data for 4000 cases on 56 items crossed with two background characteristics that were known – school type and socio-economic status. Response data was simulated using the 2-PL IRT model with items calibrated using marginal maximum likelihood estimation procedures. Their results showed that when the booklets had fewer items, person ability estimates became less accurate; and differences that existed between sub-groups became underestimated when these existed. Moreover, as test participants increased, recovery precision for the item difficulties increased. Yet, the gain in precision was more for the difficult items than was for the easier items.

Importantly, even though a dearth of literature exists investigating parameter recovery with the use of multi-matrix booklet designs, most of the few studies carried out use completely simulated data. This raises questions about the generalizability of the results to real life test data, especially as simulated data often fit perfectly to underlying IRT models used.

1.4 Aims and scope

As discussed above, large-scale educational assessments are becoming indispensable, and multi-matrix designs in combination with item response models form the state-of-the-art methodology in such assessments. A key objective in these large-scale assessments is

[4] The different booklet designs were created in the same way as in von Davier, Gonzalez & Mislevy (2009) in which they showed the adverse effects of not using plausible values correctly when analysing LSA data.

accurate estimation of population and item parameters (Beaton & Barone, 2017). Accurate estimation of these parameters is not only required at the global population level, but, importantly, also required at the level of subgroups within the population. This is crucial for educational policy making, since it provides accurate information about performance differences between population subgroups or subpopulations, which can thus guide evidence-based educational interventions (Seatrom, 2017). Importantly, although multi-matrix designs in conjunction with IRT remain state-of-the art methodology applied in these large scale educational studies, relatively less research has been carried out on these designs, with most information on them "relegated to the pages of technical manuals" (Rutkowski, Gonzalez, von Davier & Zhou, 2014, p.76). Further, factors such as test length, sample size, and item-person match have been found to relate with the precision with which person and item parameters are recovered in an IRT context (Finch & Edwards, 2015; Gershon, 1992; Svetina et al, 2003, Wollack et al, 2006).

This notwithstanding, a dearth of literature exists on how the above factors relate to the precision with which person and item parameters are recovered when using multi-matrix booklet designs. For instance, it is interesting to know the minimum sample size and test length requirements for obtaining accurate parameter estimates when using different multi-matrix booklet designs. Similarly, it can be interesting to investigate the extent to which item-person match relates to the precision with which person and item parameters are recovered when using various multi-matrix booklet designs.

Hence, using different multiple matrix booklet designs, this PhD project seeks to answer the following broad research questions: (1) How do test length or sample size influence parameter recovery precision (a) at the global population level and (b) concerning the mean performance difference between various policy-relevant subgroups? (2) Considering test length or sample size, how does the degree to which person and item parameter distributions match with each other affect parameter recovery precision?

These research questions are investigated under two large studies. Study 1 (an empirical study) seeks to answer the first research question, while Study 2 (a simulation study) tackles the last research question. In Study 1, real and partly simulated data are used. The partly

simulated data is used because some of the experimental conditions need a data structure that is not satisfied by the real dataset (e.g., the requirement for longer tests). However, the partly simulated data are got from the real dataset, hence preserving some of its original characteristics.

On the other hand, real data is used to simulate data for Study 2—since it is not possible to get real data where match between item and population parameter distributions is experimentally manipulated. Further, although the study of Svetina et al. (2013) mentioned earlier, also used simulated data, this study complements their study as it uses large samples with sparse data. In the former study, small samples with complete data were used; hence effects that were considered negligible with complete data, could now turn significant with sparse data.

In both Study one and study two, the balanced incomplete block design is considered especially as variants of it are currently used in many large-scale assessments (e.g. in PISA and TIMSS).

In the next three chapters (chapters 2 - 4), the conceptual and theoretical framework of the dissertation are further developed. Chapter two describes the Rasch item response theory model, especially as a multi-dimensional Rasch model was applied, for scaling item and person parameters in this dissertation project. Chapter three describes various techniques used in estimating person and item parameters in large-scale assessments. Emphasis is given to Maximum likelihood estimation and plausible value imputation, which are commonly used in large-scale assessments. Chapter four discusses the concept of missing data in planned sampling plans (multi-matrix booklet designs being good examples of planned sampling plans). Chapter five and Chapter six each present one of the two large studies carried out in this dissertation project (Study 1 the empirical study, and Study 2 the simulation study). The dissertation closes with chapter seven which is a general discussion of the entire doctoral project.

Chapter 2 The Rasch Item response Theory Model

The studies carried out in this dissertation used the Mixed Coefficient Multinomial Logit Model (MCMLM) for scaling item and person parameters. The MCMLM is a multi-dimensional Rasch IRT model. This chapter thus briefly introduces item response theory, spells out some of the very important assumptions of IRT models; and, concludes with a discussion on the dichotomous Rasch model (giving special emphasis to the MCMLM which is a multidimensional form of the dichotomous Rasch model).

2.1 Introduction to item response theory

After several years of slow and unsystematic growth, item response theory has grown into a fully developed and robust substitute to the classical theory of test scoring and item analysis (Bock, 1997). Attention first became drawn on such a measurement framework in the 1970's when standardized tests like the Scholastic Aptitude test applied it in their development (Polit & Yang, 2014). IRT eventually turned out to be the key psychometric method in scale validation since it offered pragmatic solutions to several measurement challenges met in the construction of tests or scales (Samejima, 1969). In IRT, item parameters in a scale are estimated based on a model where persons' latent ability levels, on the measured construct, are separated from their responses to scale items (Yang & Kao, 2014). A monotonically increasing function is used to express the relationship between a person's response pattern (to a set of test items) and their ability level (Price, 2017).

On the other hand, Classical Test Theory—the older and more traditional approach in the field of education—gives results which depend greatly on the test and sample used

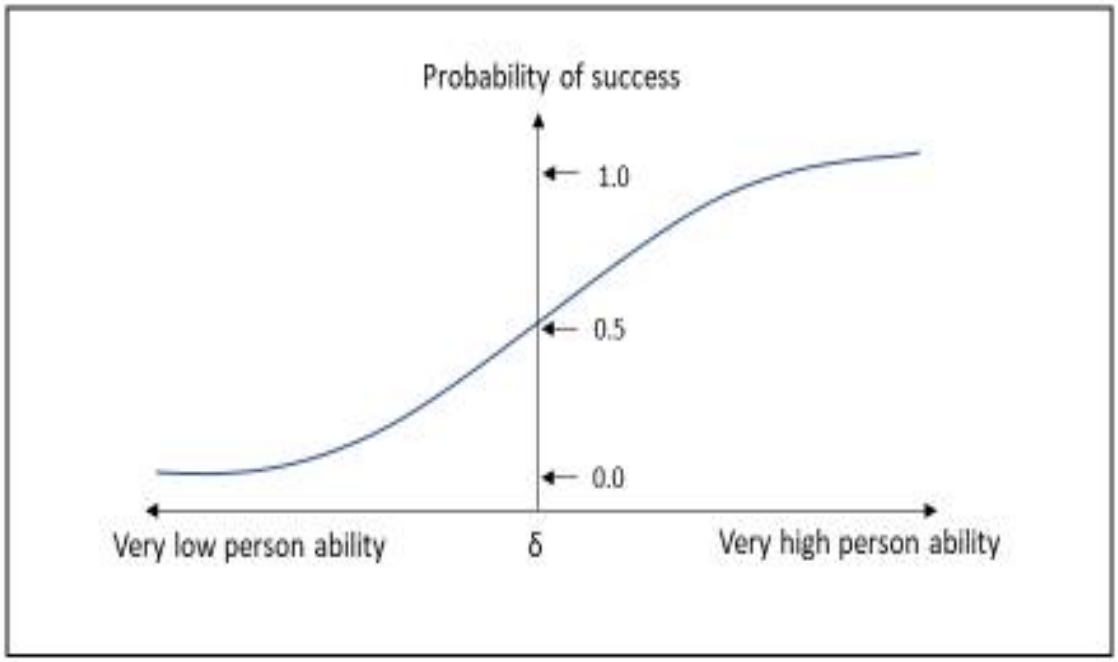
(Embretson & Reise, 2000). In this measurement framework, the raw score stands for the total responses of a person to a test or scale—signifying the person's average score given they had responded to the test an infinite number of times—plus, a random error of the summed scores from the test items (Yang & Kao, 2014). Further, since it is not possible to respond to an item an infinite number of times, this raw score can be considered a hypothetical measure of ability. Such CTT tests were often used in situations where the sample of persons taking the test had characteristics like those of persons used during initial test development (De Ayala, 2009; Yang & Kao, 2014). A major disadvantage of this is that, if the test items are difficult, calculated person abilities will tend to be low; while the same persons will have high abilities when easier items are used. For this reason, IRT was developed in which test characteristics depend solely on the characteristics of the test and not on the sample used (Yang & Kao, 2014). Consequently, estimates of item parameters remain constant irrespective of the group to which these items are administered and likewise estimates of person parameters remain constant irrespective of the tested group (Toland, 2008). This remarkable property of IRT models is referred to as the invariance property and is considered the keystone of IRT (Embretson & Reise, 2000; Hambleton et al, 1991; Lord, 1980). A further advantage of this is that it allows the construction of tests through judicious choice of items to derive very precise measurement for individuals taking the test (as in computerized adaptive testing, CAT); and offering mechanisms for putting together different tests on the same scale as in tests linking and scaling (Carlson & von Davier, 2013).

2.1.1 The dichotomous Rasch Model

Several IRT models exist. However, the model with the simplest specification is the dichotomous Rasch model (with each item scored as either correct or wrong, 0 or 1). This model was first proposed by the Danish statistician Georg Rasch for educational tests, at the same time with related models which he called models for measurement (Rasch, 1960; Kreiner, 2013). Since then, the Rasch model has grown to include several other statistical models.

IRT models use mathematical functions in modelling probabilities of students responding to test items. Graphs displaying these probability functions typically have an S-shape and are called item characteristic curves, ICC (Baker & Kim, 2017; Wu, Tam, & Jen, 2016).

Fig. 2.1. An example of an item characteristic curve



Note. Figure adapted from *Educational Measurement for Applied Researchers: Theory into Practice*, p. 95 by M. Wu, H. P. Tam and T.-H. Jen, 2016, Singapore: Springer. Copyright 2016 by Springer Nature Singapore Pte Ltd.

Consider $X_{ni} = x \in \{0,1\}$ is a dichotomous random variable with, for instance, $x = 0$ signifying an incorrect response and $x = 1$ signifying a correct response to a given test item. Using the Rasch model for dichotomous data, the probability of the outcome $X_{ni} = 1$ will be given by:

$$\Pr\{X_{ni} = 1\} = \frac{e^{\theta_n - \delta_i}}{1 + e^{\theta_n - \delta_i}}, \qquad (2.1)$$

where θ_n is the ability of person n and δ_i is the difficulty of item i (Wu, Tam, & Jen, 2016). $\Pr\{X_{ni} = 1\}$ thus, denotes the probability of a given test taker succeeding at a given item. The

Rasch model is often referred to as the one parameter (1PL) model since only one item parameter, the item difficulty (or delta, δ, parameter) is used in describing person ability.

Further, several psychological tests are based on the Rasch model, some of the most popular examples including (Kubinger & Draxler, 2007, p.294): The BAS II (British Ability Scales II; Elliot et al, 1996) and its American version DAS (Differential Ability Scales; Elliot, 1990); the K-ABC (Kaufmann Assessment Battery for Children; Kaufmann & Kaufmann, 1983); and, the AID2 (Adaptive Intelligence Diagnosticum – v 2.1; Kubinger & Wurst, 2000) within nations speaking German.

The preference of the Rasch model over other alternative IRT models is mainly for two major reasons. First, the model is very simple, making use of just one item parameter to describe a test taker's proficiency. This results—when the Rasch model fits to the data—in accurate parameter estimation with the use of fewer items than in other IRT models (Birnbaum, 1968). Second, the Rasch model has a very useful characteristic in that, examinee observed item scores can be summed up to represent an adequate statistic: This property is called sufficiency of the unweighted raw score (Fischer, 1995). This gives a fair and sufficient description of the empirical association between proficiencies of examinees who took the test and those who did not, on condition that the dichotomous Rasch model (or a monotone transformation of it) is true for the set of items under consideration (Kubinger & Draxler, 2007).

The Rasch measurement model is one out of a group of models that could be used to model data, reflecting the structure of the observations. Depending on the nature of item responses or assumptions of the composition of the total population (e.g., one general population vs. a mixture of several latent subpopulations) a suitable model can be chosen from the family of Rasch models (Tolonen, 2005). This family of Rasch models include the original dichotomous Rasch model (Rasch, 1960/1980), the Rating scale model (Andrich, 1978; Wright & Masters, 1982), the partial credit model (Masters, 1982; Wright and Masters, 1982) an extension of the rating scale model; and later several developments like the facets model (Linacre, 1989) and the Saltus model (Wilson, 1989; Draney, Wilson, Glück, & Spiel, 2008). Distinguishing characteristics of the Rasch family of models include separable item and

person parameters, sufficient statistics, and conjoint additivity—since item and person parameters can be concatenated (Masters & Wright, 1997, p. 101).

On the other hand, the original concept of the dichotomous Rasch model is expanded and modified to include a large family of other Rasch models. Wright and Mok (2004) give a description of four of these additional Rasch models: binomial trials, Poisson counts, rating scale models, and partial credit models. Binomial counts involve a situation in which an individual makes several independent trails at an item and the total number of successes recorded; however, when the number of trials gets infinitely large and the success probability very small, the binomial distribution approximates the Poisson distribution (Wright & Master, 1982, Wright & Mok, 2004). However, the rating scale model is a special case of a polytomous model, that is, a model having more than two response categories like , “strongly agree”, “agree”, “disagree”, “strongly disagree” in a four-category Likert-type scale (Wu & Adams, 2007). Further, threshold parameters are added to describe the relative difficulty of changing from one category of the rating scale to the other (Eckes, 2015). The partial credit model differs from it (i.e., the rating scale model) in that the threshold parameters are different for each item and the model is most suitable when test items contain different number of response categories for the items; or, when the relative difficulty between response categories could change from one item to the other (Masters, 1982, 2010; Wright & Mok, 2004).

2.1.2 Mixed coefficient multinomial logit model (MCMLM)

A requirement for IRT models used in most current applications is that the tests be unidimensional (Kang, 2006). However, most psychological and educational tests are to some degree multidimensional (Ackerman, 1994; Luecht & Miller, 1992; Reckase, 1979, 1997; Traub, 1983). There is thus a need to inspect test dimensionality when applying IRT models (De Ayala & Hertzog, 1991); and to apply multidimensional item response theory (MIRT) models when it is necessary to take into consideration such observed multidimensionality (Crichton, 2016).

Simply defined, MIRT models are generalizations of unidimensional models with the inclusion of additional trait or ability parameters, with the multiple traits or abilities oftentimes matched with specific problem types (Kang, 2006). An example could be when evaluating students' performance on a mathematics problem that allows for multiple solution strategies; with the possibility of a lower arithmetic knowledge being compensated for by stronger geometric knowledge. MIRT models therefore offer a perfect basis for modelling performance in complex domains, while considering multiple basic abilities at the same time and showing various ability mixtures needed for different test items (Hartig & Höhler, 2009).

Ever since the Rasch model (Rasch, 1960) was put forward, numerous extensions and alternatives have surfaced. The proliferation of these models, in some ways, has hindered test practitioners, as oftentimes, each model has its own parameter estimation techniques and dedicated software programs (Adams, Wilson & Wang, 1997). The MCMLM (Adams, Wilson & Wang, 1997; Adams & Wu, 2007) bridges this gap by offering a generalized item response model, providing a unified framework for a large class of Rasch-type models. Benefits of a single framework include mathematical elegance, generality in a single software package, and facilitation in developing, testing, and comparing new models (Adams & Wu, 2007). This model (the MCMLM) contains mixed coefficients, with items characterized by a fixed set of unknown parameters ξ, and student outcome levels (the latent variable), θ, considered a random effect (Monseur & Adams, 2009). Also, because of its several advantages, the MCMLM is used in the scaling of PISA data (OECD, 2012).

In OECD (2012, p.129-130) a detailed description of the MCMLM model is presented, which is summarized (adopting the notation of OECD, 2012) below as follows:

> Assuming we have $I = 1, \ldots, I$ items and $k = 0, \ldots, K_i$ possible response categories per item.
>
> Consider $X_i = \left(X_{i1}, X_{i2}, ..., X_{iK_i}\right)^T$ and

$$X_{ij} = \begin{cases} 1, & \text{if response to item } i \text{ is in category } j \\ 0, & \text{otherwise} \end{cases} \qquad (2.21)$$

The vector with zeros (that is, the zero category) serves as the reference category, and is required to identify the model. The probability of responding in category *j* to an item *i* is considered to have the form:

$$\Pr(X_{ij} = \mathrm{l}; \mathrm{A}, \mathrm{B}, \xi \mid \theta) = \frac{\exp(b_{ij}\theta + a_{ij}^{'}\xi)}{\sum_{k=1}^{K_i} \exp(b_{ij}\theta + a_{ij}^{'}\xi)}, \qquad (2.22)$$

Where the vector $\xi^T = (\xi_1, ..., \xi_p)$ represents the items, with ξ describing the empirical characteristics of the response categories per item; $\mathrm{A}^T = (a_{11}, a_{12}, ..., a_{1K_1}, a_{21}, ..., a_{IK_I})$ is a design matrix created from a set of design vectors, $a_{ij} (i = 1, ..., I; j = 1, ..., K_i)$, $\mathrm{B} = (\mathrm{B}_1^T, \mathrm{B}_2^T, ..., \mathrm{B}_I^T)^T$ is the scoring matrix sub-matrix for item *i*, and $b_{ik} = (b_{ik1}, b_{ik2}, ..., b_{iKD})^T$ a vector consisting of scores across D dimensions, and $\theta = (\theta_1, \theta_2, ..., \theta_D)^{'}$ describes the position of an individual in the D-dimensional latent space .

The response vector is given by:

$$f(x; \xi \mid \theta) = \psi(\theta, \xi) \exp(x'(\mathrm{B}\theta + \mathrm{A}\xi)), \qquad (2.23)$$

$$\psi(\theta, \xi) = \left(\sum_{z \in \Omega} \exp(z^T (\mathrm{B}\theta + \mathrm{A}\xi)) \right)^{-1} \qquad (2.24)$$

where Ω represents the set of all response vectors possible and *x* a given case of X.

Due to the several advantages of the MCMLM (as described above), and its popularity in many large-scale assessments, for example in PISA; this model was applied throughout this dissertation project, for the scaling of person and item parameters.

Chapter 3 Estimation of person and item parameters

Several techniques and procedures exist for estimating person and item parameters within an IRT framework. Si and Schumacher (2004, p. 154) list several of these techniques, which include: maximum likelihood method (Baker, 1992); logistic regression (Reynolds, Pekins and Brutten, 1994); Bayesian modal estimation (Mislevy, 1986; Baker, 1992); and the minimum chi-quadrant estimation technique (Zwinderman & van der Wollenberg, 1990). I will emphasize on the maximum likelihood procedure since this was utilized in this dissertation project. This will be followed by a description of the plausible values methodology, which is state-of-the-art methodology used in large-scale assessments, for estimating population statistics of student proficiencies. The chapter will conclude with a discussion on efficiency measurement in IRT.

3.1 Maximum likelihood method

Maximum likelihood is the second most widely used missing data treatment method, after multiple imputation, with many modern statistical techniques largely depending on it (Baraldi & Enders, 2010; Enders, 2004; Schafer & Graham, 2002). The basic principle behind maximum likelihood involves choosing estimates that maximize the probability of getting the results that are observed (Allison, 2002). This is done by identifying population parameter values with the greatest likelihood of producing the sample data, using a mathematical function—known as the log likelihood—to quantify the standardized distance between observed data points and parameters of interest (Baraldi & Enders, 2010).

Baraldi & Enders (2010, p.19) summarized the basic principle of maximum likelihood estimation as follows:

- Using a mathematical function known as the log likelihood to quantify standardized distances between observed data points and parameters of interest (e.g. means).
- For a sample of scores, this log likelihood is given by

$$\log L = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2} \right] \qquad (3.1)$$

  Where the term in brackets represents the probability density function and describes the shape of the normal curve.

- In combination, the term in brackets in (3.1) specifies the relative probability of getting a unique score with a given mean and standard deviation from a normally distributed population; with the summation symbol including relative probabilities into the sample log likelihood (a summary measure quantifying the likelihood of choosing the entire sample from a population that is normally distributed).
- Using an iterative algorithm to repeatedly substitute various parameter values into the log likelihood equation to obtain the highest value or probability (in other words, continuing with the iterative process until estimates that minimize the distance to the data are obtained).

Three kinds of maximum likelihood estimations are often applied in IRT parameter estimation. These include Joint Maximum Likelihood, JML (Birnbaum, 1968); Conditional Maximum Likelihood, CML (Andersen, 1972); and, Marginal Maximum Likelihood, MML (Bock & Liberman, 1970). In JML, item and person ability parameters are estimated simultaneously using a two-staged iterative procedure. This is done by treating both person and item parameters as unknown but fixed model parameters and estimating them together by solving an optimization problem (Chen, Li, & Zhang, 2018). Although a major advantage of JML estimation is its simplicity and straightforwardness, it yields inconsistent, estimated

parameters for fixed length tests, with person parameters being infinite when no or all test items are correctly endorsed (Embretson & Reise, 2000). It is thus not suitable with items or examiners having perfect scores and cannot be used in small scale studies; as long tests with large samples are needed to minimize parameter estimation bias (Le & Adams, 2013).

In CML, item and person ability parameters are estimated separately by conditioning the likelihood function on examinee ability (Si & Schumacher, 2004). In this estimation technique, maximization of the log-likelihood function is simplified by treating some of the parameters as if they are known—with these parameters either fixed by theoretical assumptions or, more often, replaced by estimates (Palmer, 2011). However, CML is unable to compute parameter estimates for perfect or zero scores and can assign different scores to examinees having the same number of correct responses (Si & Schumacher, 2004). Also, although CML estimates for item difficulties are consistent, an ad hoc technique ought to be implemented to estimate person abilities—thus, CML is appropriate only when a simple sufficient statistic such as a raw score for a Rasch model is available (Johnson, 2007).

Conversely, while CML uses the likelihood function conditioned on examinee ability, MML utilizes the unconditional likelihood function—which is the probability of obtaining a given pattern of scores from an examinee with unknown ability, randomly chosen from the population (Si & Schumacher, 2004). MML thus handles the problem of unknown person abilities by using the unconditional likelihood function instead of the conditional likelihood (as in CML). Unlike in JML where person parameters are treated as fixed effect parameters and kept in the likelihood function, MML treats person parameters as random effects and integrates them out from the likelihood function (Chen, Li, & Zhang, 2018; Si, 2002).

Importantly, Bock and Lieberman (1970) used a general procedure which involved maximizing the likelihood in the marginal distribution, after having performed a numerical integration over the latent distribution. However, the computational complexity of the estimation technique made it impractical for lengthy tests (Baker & Kim, 2004; Bock & Moustaki, 2007; Si & Schumacher, 2004). Thus, Bock & Aitkin (1981) proposed a feasible approach for estimating the item parameters in large scale tests by applying a reformulation of the Expectation-Maximization (EM) algorithm (Chalmers, 2012; Dempster, Laird, &

Rubin, 1977; Si & Schumacher, 2004). The EM algorithm consists of two steps—The Expectation (E) and the Maximization (M) steps. At the E stage, the expected score function of model parameters are calculated—with expectations being with respect to posterior distributions of given observations; while at the M stage, obtained parameters are updated by applying them in the marginal likelihood estimation equations (Bock & Moustaki, 2007). The expectation and the maximization phases are rerun severally until convergence of the estimates with the maximum likelihood equation occurs—this achieved by applying the Newton-Gauss procedure to solve the equations (Si & Schumacher, 2004).

However, MML estimation has some short comings. For instance, the technique is not suitable for analyses in non-regular distributions—where a maximum likelihood may be non-existent, or several maximum likelihoods present, thus invalidating the idea of maximizing the likelihood (Cousineau & Allan, 2015). Further, the technique is computationally intensive and requires an assumption being made about the nature of the ability distribution—if this is initially not ascertained, a normal distribution is often assumed (Si & Schumacher, 2004). Lastly, though using the EM algorithm remedies the issue of unstable item parameters with JML estimation, an unsolved problem remains aberrant ability estimates when using certain datasets, with parameter estimation for irregular response patterns being impossible (Baker, 1992; Si & Schumacher, 2004).

Despite the above short comings, MML still remains a technique highly recommended by methodologists for the estimation of item and population parameters in IRT (Toland, 2008). First, it gives consistent estimates of item parameters irrespective of the sample size, as greater sample sizes do not demand more examinee parameters to be estimated (Baker & Kim, 2004). Second, MML provides item standard error estimates which efficiently approximate the expected sampling variance, and can be used to compute ability and item parameter estimates for test takers with perfect or zero scores; hence no information loss due to deleting items and persons with such scores (Si & Schumacher, 2004). Third, MML solves some of the problems with the JML method by introducing an assumption on the latent variable distribution (Le & Adams, 2013). Lastly, the estimation technique is currently well-implemented in most popular statistical software packages, thus reducing challenges due to

its computational complexity. Due to these numerous advantages, MML estimation was used in this dissertation for the estimation of item and population parameters.

3.2 Plausible values (PVs) imputation approach

Large scale assessments often face the challenge of missing data values, since multiple matrix sampling is used in their design (Frey et al., 2009; Gonzalez & Rutkowski, 2010). These missing values make the uncertainty related to individual θ estimates to be large, resulting in seriously biased population estimates when individual scores are aggregated (Wingersky, Kaplan, & Beaton, 1987). Plausible values are a range of reasonable abilities a test taker can obtain given his/her responses to the test items and are randomly drawn from an estimated distribution of the test taker's ability on the measured latent trait (Wu, 2005). The distribution from which abilities are drawn for a given test taker is called the posterior distribution (Mislevy, Beaton, Kaplan, & Sheehan, 1992).

Plausible value methodology (Rubin, 1987; Mislevy, 1991) was introduced to solve the problem in which sets of scores (known as plausible values) are generated using students' responses to all items and conditioned on available background data (Yamamoto & Kulick, 2000). Conditioning on all background data ensures that relationships between background variables and the estimated person abilities are correctly accounted for in the PVs (Mislevy et al., 1992). Typical examples of LSAs in which PVs are used are PISA and TIMMS (see OEDC, 2012; Yamamoto & Kulick, 2000).

Adopting the mathematical notation and discussion of the PV methodology from Laukaityte & Wiberg (2017, p.11344-11345), this method is summarised as follows:

> Let y_i represent student i's responses to background questions, and x_i student i's item responses. Given that for each student i, 5 PVs are drawn from the conditional ability distribution then,

$$(\theta_i \mid x_i, y_i, \Gamma, \Sigma) \propto P(x_i \mid \theta_i, y_i, \Gamma, \Sigma) P(\theta_i \mid y_i, \Gamma, \Sigma) = P(x_i \mid \theta_i), \Gamma, \Sigma) P(\theta_i \mid y_i, \Gamma, \Sigma) \quad (3.2)$$

with $P(x_i \mid \theta_i)$ being any chosen response model, $P(\theta_i \mid y_i, \Gamma, \Sigma)$ the regression of the background variables, Γ a matrix of regression coefficients for the background variables, and Σ a variance matrix of residuals (Laukaityte & Wiberg, 2017). This results in a set of drawn values, $\hat{D}_m$, with $m = 1,...,M$ and $M > 1$ denoting the number of drawn PVs. Hence, the analysis is performed for every $\hat{D}_m$, with the final estimate got by taking the average of all *M* estimates:

$$\bar{D} = \frac{\sum_m \hat{D}_m}{M} \qquad (3.3)$$

The total variance of $\bar{D}$ is computed by adding up the within imputation variance and the between imputation-variance. The within imputation variance is got by taking the average of the estimated variances V_m of PVs $\hat{D}_m$, $\bar{V} = \frac{1}{M}\sum_m V_m$, and the between-imputation variance, $Var(\bar{D}) = \bar{V} + \left(1 + \frac{1}{M}\right) B_M \qquad (3.4)$

where $B_M = \frac{1}{M-1}\sum_m (\hat{D}_m - \bar{D}_m)^2$ (Mislevy, 1991; Schafer, 1997). It is noteworthy that, the PV methodology produces consistent[5] population parameter estimates, provided the PVs are generated with an imputation model compatible with data analyses that follows (Laukaityte & Wiberg, 2017).

Conversely, recent publications have raised concerns about the modelling approach used to generate ability measures in large-scale assessments like PISA and NAEP; and, whether or how to use PVs in secondary analyses (Braun & von Davier, 2017). Although in the past three decades, extensive research has been carried out on the fundamental principles and statistical methodology applied in these models (Mislevy 1984, 1985; Mislevy & Sheehan, 1987); concerns continue to arise if the resulting PVs produce appropriate estimates of population estimates like means and variances (e.g., Goldstein, 2004; Cohen & Jiang, 1999).

[5] In Statistics, an estimator is said to be consistent if, the values of this estimator become closer to the true parameter value as the sample size is increased.

Further, Jacob & Rothstein (2016) question the suitability of using latent regression methodology and PVs to produce achievement scores, later used as inputs in secondary analyses in econometric modelling. Braun and von Davier (2017) however give a detailed response to all these concerns. They present a broad review of key literature, with emphasis on important journal articles describing the psychometric properties and derivations of PV values. A simulation study is then performed which compares statistical characteristics of estimated values derived using PVs with those derived using other often used methods. Their results show that PV methodology outperforms the other methods; and produces estimates of model parameters that are approximately unbiased when using them in regression analyses. Hence, PV methodology applied in reporting examinee performance in large-scale assessments, remains state-of-the-art for individuals performing secondary analyses from such databases (Braun & von Davier, 2017).

3.3 Efficiency measurement based on item response theory

For an examinee to be measured most effectively, the administered test items need not be too easy or difficult (Lord, 1980). The implication is that, ideally, for a student population with differing abilities, different item sets of varying difficulties or different test booklets need to be utilized to efficiently estimate each test taker's ability (Braun & von Davier, 2017; Weiss, 1982). That notwithstanding, though students respond to different items—as found in either an easy, average or even a difficult test booklet—the final test scores still need to be directly comparable (Berger, Verschoor, Eggen & Moser, 2019). By applying several test equating strategies (Kolen and Brennan, 2014), item response theory provides a powerful measurement framework for achieving this goal.

Taking the simplest unidirectional IRT model, the Rasch model, an examinee's likelihood to correctly respond to a specific item is defined by:

$$P(X_{ij} = 1 \mid \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} = p_{ij} \qquad (3.6)$$

where θ_i denotes the ability of examinee *i*, and β_j denotes item *j*'s difficulty (Rasch, 1960; Rost, 2004). Further, using maximum likelihood procedures, examinee abilities and their standard errors can be estimated. According to Rost (2004), the standard error of the estimated ability for examinee *i* is given by

$$SE(\hat{\theta}_i) \sim \sqrt{\frac{1}{\sum_{j=1}^{k} p_{ij}(1-p_{ij})}} \qquad (3.7)$$

with p_{ij} denoting the likelihood that examinee *i* responds to item *j* correctly, as described in equation (3.6). Further, judging from equation (3.7), it can be inferred that (a) *k*, which represents the total number of test items administered per examinee, plays a crucial role in how accurately an examinee's ability is estimated when using the Rasch model; (b) the accuracy of estimated examinee abilities also depend on the relationship between an examinee's ability and the difficulty of the administered items in the test (Lord, 1980; Rost, 2004).

In operational testing scenarios, test length (i.e., number of test items administered per student) is often determined in advance considering available testing time. The main option left for enhancing estimation of examinee ability, and of course test efficiency, is optimising the relationship between examinee ability and item difficulty (Berger et al., 2019). Thus, resulting in the concept of targeted testing.

Targeted testing is a test construction technique where administration of test items is done to match examinee abilities, thus improving measurement efficiency. Further, choosing a suitable test design demands knowing the test purpose, and population of test takers. A test meant to classify examinees or a test targeting a specific population, demands measuring examinee ability most precisely around particular points along the ability continuum (Berger et al., 2019). A suitable test design will thus incorporate items that give large amounts of information, at the specific points that are of interest. Conversely, tests aimed at assessing student abilities in very diverse populations—like in formative assessments—demand test designs, which give results that are accurate over a broad range of student abilities. Usually, in such instances, it is inappropriate to use a single linear test with items having varying

difficulties. The reason being that typically, individual students are administered many questions that do not match their ability levels. This could thus result in reduced measurement efficiency, and also decreased student motivation during the test (Dong & Peng, 2013; Lord, 1980).

Generally, there are two ways of taking into consideration a wide variation in student ability through the targeted administration of items having varying difficulties (e.g., Mislevy and Wu, 1996). First, one could use information known in advance about examinees abilities to assign them to matching test forms. In school settings, it is often possible to get such preliminary or advance information from other similar tests, which could thus be used in assigning them to test forms that match their ability. Also, teachers assess their students in various tests and assign them to different school grades and sometimes, even to different school types or performance groups (Berger et al., 2019). Such information could be utilised to create ability groups into which examinees are distributed based on their ability. The problem with this approach though is that these background variables, which relate to student ability only approximate students' true abilities. Thus, some students could greatly differ from the group mean, and consequently, from the target ability of the test (Bejar, 2014).

Conversely, a step-by-step approach could be used to assign target items or item sets to students, based on how they perform in the course of the test. Thus, students who perform well are automatically administered more difficult test questions—allowing their full potential to be shown—while those who perform poorly automatically receive easier test items. This is the fundamental idea applied in targeted testing designs like computer adaptive testing and multi-stage testing. This concept of matching items to persons is further examined in one of the studies carried out in this dissertation project (See Chapter 6).

Chapter 4 Missing data in planned sampling plans

Missing data are a common problem in quantitative research studies (Peugh & Enders, 2004); posing serious implications which could lead to biased parameter estimates, information loss, diminished statistical power, larger standard errors, or weaker generalizability of findings (Don & Peng, 2013). This became more remarkable within the last decades, with increased availability of data from large-scale assessments—where missing values occur inevitably (Pohl & Aßmann, 2015). This chapter reviews important concepts related to missing data as pertains to quantitative research in Education and Psychology. It begins with a description of several planned missing sampling designs, followed by an explanation of the theory behind the missingness in these sampling designs. The chapter concludes with an elaboration on multiple imputation, an important technique for treating missing data.

4.1 Planned sampling plans

Planned sampling plans have enjoyed a recent growth in popularity and involve researchers deliberately collecting only partial data (Wood, Matthews & Pellowski, 2018). Also, although the resulting missing data could be considered a challenge for applied researchers, the reverse could rather be true—for instance, when the degree of missing data on a particular variable is carefully controlled, a balance can be struck between statistical power and research costs (Rhemtulla & Hancock, 2016). In the early days of missing data analyses, missing observations were treated with some form of deletion or simple imputation; however, nowadays, sophisticated options (like multiple imputation and maximum likelihood) exist for analysing datasets with high levels of missingness, provided certain conditions are met (Enders, 2010; Silvia, Kwapil, Walsh & Myin-Bermeys, 2014).

However, a major challenge in creating these planned sampling plans is ensuring that the required amount of information is still successfully collected, and that valid and reliable statistical inferences on quantities of interest are derived from these partial datasets. Thus, several possibilities exist in the way these planned sampling plans can be created, some of the important examples including item sampling, item-examinee sampling, and survey questionnaire sampling.

4.1.1 Item Sampling

This is one of the oldest and simplest test abridgment methods and involves administering a portion of the test items to all test takers (Moy & Barcikowski, 1974). As an example, for a test of 80 items, only the first 40 test items are administered to all test takers. Further, this sampling method is usually applied in the abridgment of already existing test instruments, while ensuring optimal psychometric properties for items (Coste, Guillemin, Pouchot, and Fermanian, 1997). It was thus first applied in creating new Wechsler subtest combinations such as short forms and factor scales, in which Composite Measurement Scales (CMS) were shortened by eliminating some of the test items (Tellagen & Briggs, 1967; Wolf, 2006).

According to Wolf (2006, p. 53), the advantages and disadvantages of this sampling method can be summarised as follows. For the advantages:

- The method is very easy to administer (with every test taker responding to the same test items).
- It offers a shorter test version of the same, or even higher validity and reliability.
- It provides maximum statistical power for the chosen test items (though this might not be true for the entire measurement scale).
- The method uniquely makes it possible to compute and compare scores for subjects on characteristics of interest.

For the disadvantages:

- The method is only suitable in rare instances, like where the original test or survey instrument serves as the reference for measuring the characteristics or traits of interest.
- Further, this technique is realistic only with short scales having well-defined psychometric properties, since item selection is applied per scale.

However, it is noteworthy that in older literature (e.g. Lord, 1962), the term *item-sampling* is used as an umbrella term to represent more specific sampling types like matrix sampling (Gonzalez & Rutkowski, 2010). These other sampling plans are presented below.

4.1.2 Item-Examinee Sampling

This approach involves randomly selecting items from the item universe and administering them to randomly selected subjects from the population. It is also referred to as multiple matrix sampling (MMS). Item-examinee sampling is not only used to shorten tests, but also to widen topical breath as it allows many more items to be included in the sample of administered test items (Gonzalez & Rutkowski, 2010). It is well-suited when estimating group or sub-population measures, but unsatisfactory for individual diagnostics. In classical non-overlapping item-examinee sampling, theoretical characteristics of the sampling distribution are unknown, making hypothesis testing or the creation of confidence intervals impossible (Thomas, Raghunathan, Schenker, Katzoff & Johnson, 2006). Thus, in modern MMS, information about inter-item relationships is integrated into parameter estimation procedures, resulting in significant efficiency increases.

In the first chapter of this dissertation, an introduction was given to multiple matrix sampling in which the popular and most-often used Balanced Incomplete Block MMS Design was discussed. Below, is a description of some additional designs that use MMS.

First, the Split Questionnaire Survey (SQS) Design was introduced by Raghunathan & Grizzle (1995). The sampling technique in this multi-matrix design makes it possible for different patterns, or sets, of data items to be collected from different sample respondents;

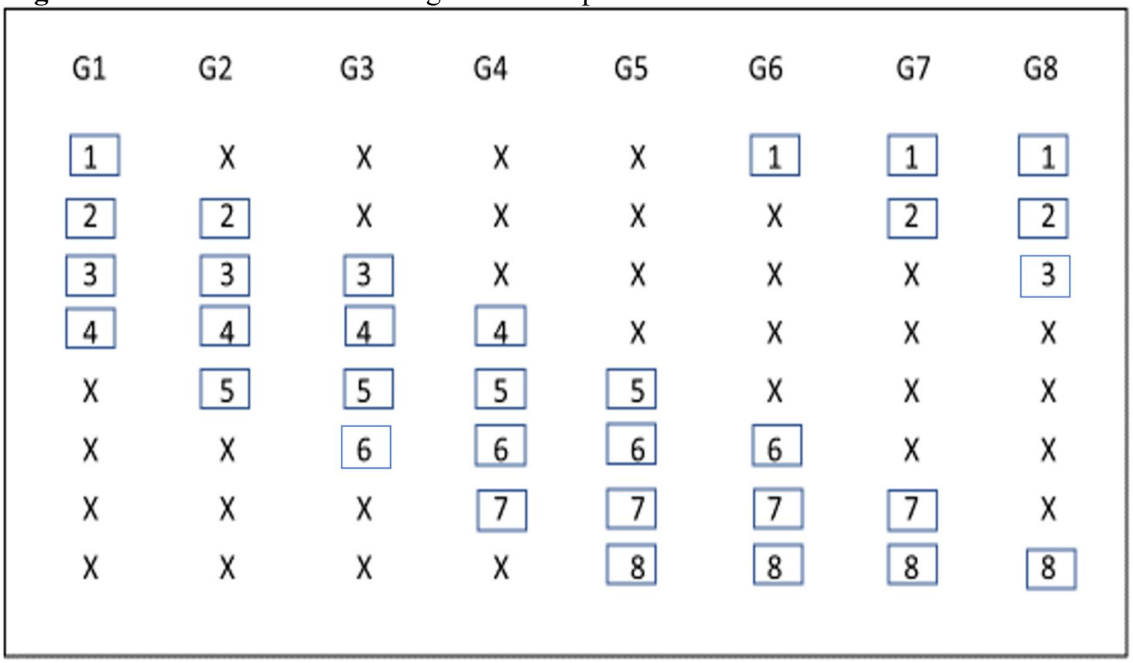
and contains item blocks which overlap though items are not randomly assigned into these blocks (Chipperfield, Barr & Steel, 2018). Here, pilot data is used to calculate partial correlations between pairs of variables, and this is used in assigning items to subsets (Rhemtulla & Hancock, 2016). This improves efficiency since items administered to the same individuals belong to the same item blocks and have weak correlations; while those not administered together (and belonging to different item blocks) have high correlations. Further, since variables measuring similar constructs tend to have high correlations with one another, the amount of information in a planned missing design is better maximized when individual participants miss only *some* items on each scale while the other items are observed (Rhemtulla & Hancock, 2016). However, there is a likelihood that some item pairs fail to occur together making it difficult to estimate their corresponding associations; thus, the design dictates restrictions to be observed when assigning items to individuals sampled for use in desired population parameter estimation (Raghunathan & Grizzle, 1995).

Second, the Three Form Design was introduced by Graham et al (1984) and has since then been used in several other studies (e.g., Graham, Johnson, Hansen, Flay, & Glee, 1990; Hansen, Johnson, Flay, Graham & Sobel, 1988; Graham, Taylor, & Cumsille, 2001). The design's aim is to reduce how long it takes to complete a survey, administer more test items than can be answered by an individual test taker, and ensure that all correlations, means and variances can be estimated (Graham, Taylor, Olchowski, & Cumsille, 2006). To do this, the test items to be used in the test are first distributed into four item sets (X, A, B, C). Using the X item set as a common set administered to all test takers, three test forms are created with each test form containing the X set of items and two other item sets from either the A, B or C item sets (i.e., Form 1 – XAB; Form 2 – XAC; Form 3 – XBC). Also, several variants of the Three Form design have been suggested. For example, XABC, XCAB and XBCA (Flay et al, 1995); X_1ABX_2C, X_1CAX_2B, X_1BCX_2A (Taylor, Graham, Palmer, & Tatterson, 1998). In the Taylor et al (1998) variant of the Three Form design, the common set of items is split into two parts – with one part administered at the beginning of the test and the other part towards the end of the test. This is important because, it helps to mitigate order effects that could result from administering the common set of items only at the beginning of the test. Also, order effects are considered in the Flay et al (1995) variant because, although the common

item set is administered at the beginning of the test in all test forms, all other three booklets (A, B, C) are used, with their positions being fully rotated across all the test forms.

Further, the fractional block design was proposed by McArdle (1994). This design improves upon simple matrix sampling making it possible to estimate means for all variables, as well as correlations for most—though not every—pair of variables (Graham et al., 2006). Considering a study collecting information on 8 variables and for a population randomly distributed into 8 groups, Figure 4.1 illustrates the implementation of a fractional block sampling design in this population (with each group measured on 4 out of 8 variables investigated in the study). The number of independent groups (G) used in this design depends on the number of measured variables and the desired spread of the various variable pairs; thus, in the above example, every variable is measured in four groups, resulting in an overall balance of sample sizes for means and standard deviations (McArdle, 1994).

Figure 4.1. A fractional block design for incomplete measurements



Note. G1, G2, …, G8 represent the different groups in the population. The squares represent the measured variables while 'X' represent missing and unmeasured variables. Figure adapted from "Structural Factor Analysis Experiments with Incomplete Data" by J. J. McArdle, 1994, *Multivariate Behavioural Research*, 22 (4), p. 424. Copyright 1994 by Lawrence Erlbaum Associates, Inc.

A disadvantage of the design is however that the overall balance does not hold for all correlations and the design demands using specialized structural equation modelling techniques (Graham et al., 2006).

4.2 Missing data theory in sampling plans

Before the 1980's partial sampling plans were clearly not described as planned missing data procedures, since inferences were only based on available data. Nowadays, missing data are replaced with imputed datasets, making it safe to consider these designs as planned missing data designs (Shin, 2016). In earlier studies, simple missing data treatment methods like pairwise deletion were used. Simulations were carried out in which person and item characteristics were manipulated to discover conditions for optimal recovery of population estimates.

However, with the emergence of sophisticated missing data techniques (Rubin, 1976, 1977) like maximum likelihood and multiple imputation, these sampling designs can now conveniently fall under the general category of *planned missingness* (Baraldi & Enders, 2010). Further, this improved comprehension of procedures resulting in missing data and the advanced methods (like maximum likelihood and multiple imputation) resulted in sampling methods which produce more efficient population estimates for tests with partial data (Baraldi & Enders, 2010). In the following sub-section, a resumé of missing data theory will be given with emphasis on missing data mechanisms and a comparison of various missing data techniques[6].

4.2.1 Missing data mechanisms

Missing data is often encountered when carrying out research in the social sciences. Missing data mechanisms explain how the measured variables are related and how likely missingness is expected to occur (Rubin, 1987, Crichton, 2016). Since various kinds of missing data

[6] An in-depth introduction and overview can be found in Allison (2001); Little & Rubin (2002); and Schafer & Graham (2002).

mechanisms underlie the missing responses in any given test or survey, it is necessary to consider this when treating missing data—as, this ensures more accurate inferences from analysis results (Crichton, 2016). The missing data mechanism differs from the missing data pattern in that the former describes why missingness took place, while the later simply describes the position of the missing data (Enders, 2010; National Research Council, 2010). As an example, holes (indicating missing data) found all around a dataset in no clear pattern, do not definitely imply that the missing data mechanism is random (Shin, 2016). Irrespective of the missing data pattern, the underlying missing data mechanism could be systematic. In this case, it implies the probability for data to be missing is linked to an underlying characteristic of the variable of missing (Crichton, 2016; Shin, 2016).

Missing data mechanisms are thus generally placed into three major categories (Crichton, 2016; Köhler, 2017; Rubin, 1987): missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In MCAR, the probability of missing does not depend on either the observed data, the value of the missing data itself (i.e., speculative data which could have been observed if there was no missing), nor on some other variable in the analysis (Enders, 2010; Shin, 2016).

The section below adopts the discussion (and mathematical notation) of missing data mechanisms (i.e., MCAR, MAR, and MNAR) from Shin (2016, p.15-16). Similar discussions can be found in Enders (2010) or Shafer and Graham (2002). Given that the complete data consists of observed and missing data, this could be described as:

$$Y_{com} = (Y_{obs}, Y_{mis}), \tag{4.1}$$

where Y_{com} denotes the complete dataset, Y_{obs} the observed data, and Y_{mis} the missing data. Therefore, MCAR can be represented by,

$$\mathrm{P}(\mathrm{M} \mid Y_{com}) = \mathrm{P}(\mathrm{M}), \tag{4.2}$$

where $\mathrm{P}(\mathrm{M} \mid Y_{com})$ denotes the probability of having missingness when using complete data, and M a matrix (of 0's and 1's) describing missingness in the dataset, "1" standing for

missing responses, and "0" for non-missing responses (Shin, 2016). Equation (4.2) shows that the probability that missingness occurs is independent of the data (Shin, 2016). It can therefore be assumed that for MCAR to hold, the observed data must be random samples from the complete dataset (Ali, Dawson, Blows, Provensano, Ellis, Baglietto, Huntsman, Caldas, & Pharoah, 2011). MCAR thus describes data in which complete cases are a random sample of the originally identified set of cases (Pigott, 2001). Unfortunately, this assumption rarely holds in real life testing scenarios since when dealing with human subjects, a high probability exists that an underlying factor could increase the response likelihood for certain subjects (Crichton, 2016).

On the other hand, data are MAR provided missingness is associated with some other measured variables in the analysis model; although such missingness is unrelated to any hypothetical values that could have been obtained given that the data were complete (Baraldi & Enders, 2010). MAR demands less strict assumptions on the reason for missingness and oftentimes occurs in practice; hence, this condition is often assumed to be true when applying most missing data techniques (Kang, 2006). In MAR, missingness and the variable of missing are independent (Enders, 2010). Hence, this missingness is thus represented by:

$$P(M \mid Y_{com}) = P(M \mid Y_{obs}), \tag{4.3}$$

where, $P(M \mid Y_{obs})$ denotes the probability of missingness, taking into consideration only the observed sections of the data (Shin, 2016). From Equation (4.3), it is evident that the probability of missingness is only related to the observed data, and not to the missing data – in other words, the probability that missingness occurs is unrelated to the underlying missing data provided the measured variables are considered (Shin, 2016). An example of MAR could be in a study estimating levels of depression in a population, with females less likely to report that they suffer from depression than males (Crichton, 2016). Thus, in this case, missingness depends partly on the sex of the participant and gender can serve as a good factor when accounting for the missingness.

Conversely, missing not at random (MNAR) arises when missingness is systematically associated with hypothetical values which are missing, implying data are missing based on

expected values of missing scores (Baraldi & Enders, 2010). Thus, this missingness can be described as,

$$P(M \mid Y_{com}) \neq P(M \mid Y_{obs}) \qquad (4.4)$$

implying that the MAR condition fails, with observed data not fully explaining missingness, and missingness rather related to the part containing missing data (Rubin, 1976; Enders, 2010; Shin, 2016). Therefore, MNAR occurs when the probability that an item will be omitted depends on hypothetical responses to the missing items after considering parts of the dataset that are observed (Mislevy & Wu, 1996; Shin, 2016). Further, when data is MNAR, there is need to have extensive prior knowledge about the missing data process, since this needs to be specifically modelled as part of the estimation process (Cheema, 2014).

Importantly, the main distinction between MAR and MCAR is if missingness is associated to the theoretical missing values. Thus, if missing values are unknown, a distinction is impossible. Nonetheless, the probability that data is MAR and not MNAR can be notably increased (Dong & Peng, 2013). For example, when test takers omit items to which the correct response is unknown, this results in MNAR because the likelihood of missingness is associated to whether an item can be answered correctly or wrongly (Dong & Peng, 2013; Shin, 2016). Further, although one might want to consider the three missing data mechanisms as mutually exclusive, it could happen that all three occur in one dataset based on which variables are included in the analysis model (Baraldi & Enders, 2010; Peugh & Enders, 2004; Yuan & Bentler, 2000).

In the context of multiple matrix sampling, examinees are administered subsets of items, resulting in these examinees having complete data on the administered item blocks and missing data on incomplete blocks. Consequently, this results in data which are MCAR, since the missing item blocks, by definition, are not related to the underlying achievement of examinees nor to other variables measured in the dataset (Peugh & Enders, 2004; Shin, 2006). Further, missing data which are MCAR or MAR can be considered ignorable and will not result in biased parameter estimates, for instance regression weights (Rubin, 1976; Enders, 2010). On the other hand, when missing data are MNAR, the mechanism controlling

missingness becomes nonignorable requiring special models that must incorporate this missingness (Howell, 2007). However, although it is expected that applying modern multiple matrix sampling like the BIBD will result in unbiased parameter estimates, standard errors and confidence intervals of estimated parameters can still be affected by the amount of missing data in these designs (Rhemtulla & Hancock, 2016). Thus, the need in this dissertation project to investigate the efficiency with which population and item parameters are recovered under several conditions in different matrix booklet designs.

4.2.2 Missing data treatments

The treatment of missing data has evolved over the years from simple conventional methods (e.g., listwise deletion, pairwise deletion, mean substitution, regression substitution, and hot deck substitution) to more complex modern methods like maximum likelihood, full information marginal likelihood, and multiple imputation (Baraldi & Enders, 2010; Shin, 2016). Listwise deletion involves discarding from a calculation (or series of calculations) such as a correlation matrix, all cases containing any amount of missing data; while pairwise deletion involves discarding information only from those statistics that "need" the information (Roth, 1994). Mean substitution replaces missing values with the arithmetic mean of available cases; regression or conditional mean imputation replaces these missing values with scores obtained from a regression equation; while hot deck imputation imputes missing values with scores from "similar" respondents in the current dataset (Enders, 2010; Wolf, 2006).

Application of the conventional missing data treatment methods was generally encouraged by their simplicity and lack of complications in their procedure. However, these conventional treatment methods rely on the stringent condition that the missing data be MCAR (a condition rarely satisfied in practice); with the deletion methods particularly resulting in data loss and thus, loss in statistical power (Dong & Peng, 2013) . Further, these methods underestimate standard errors by not considering two important sources of variance—random error resulting from the missing data; as well as, the random error of treating the imputed missing data estimates like true values (Crichton, 2016; Shin, 2016). As outlined in Allison,

2009, p.75), three generally accepted conditions for a good missing data treatment method should be:

1. It should minimize bias in the estimated parameters to the smallest possible value. This so because, it is well established that missing data can introduce bias in estimating parameters of interest.

2. The treatment method should make the greatest use of available data. Consequently, it should greatly avoid discarding some already available data. Rather, all available data should be used (if possible) so that parameter estimates with minimum sampling variability are produced.

3. The method should also produce good uncertainty estimates. Thus, estimates of standard errors, confidence intervals and p-values should be accurate.

Adding to the above, it is desirably that these conditions be achieved without making unnecessarily restrictive assumptions about the missing data mechanism (Allison, 2000). Interestingly, complex modern missing data treatment methods like multiple imputation (applied in this dissertation) perform quite well in satisfying these conditions.

Multiple imputation was introduced by Rubin (1987) and currently one of the most popular missing data treatment methods. The basic idea behind this method can be summarized as follows (Allison, 2000, p.301):

1. A suitable model is used to impute missing values, taking into consideration random variation.

2. This is done *M* times producing *M* "complete" data sets. The number of times, *M*, usually depends on the amount of missing data in the dataset, with *M* being larger for datasets with more missing data.

3. Using standard methods applied for complete data to carry out the desired analysis on each of the imputed "complete" datasets.

4. Get a single-point estimate by taking the average value of the parameter estimate of interest across the M imputed data sets.

5. Compute standard errors by (a) taking the average of squared standard errors from the M estimates, (b) calculating variance of M parameter estimates across samples, and (c) using a simple formula[7] to combine the two quantities.

A thorough discussion of this method can be found in Little and Rubin (1989); Schafer (1997); and in Schafer and Oslen (1998). Further, when performing multiple imputation, two important factors to consider include how many imputed datasets to use and which auxiliary or support variables to include in the imputation model (Shin, 2016).

Several authors recommend different guidelines as to the number of imputed datasets to use in order to obtain accurate parameter estimates. For instance, Rubin (1987) and Schafer (1997) recommend using three to ten imputed "complete" datasets; while Enders (2010) and Graham, Olchowski & Gilreath (2007) recommend using at least twenty datasets. On the other hand, White, Royston and Wood (2010) recommend basing the number of imputed datasets on the percentage of missing data in the original dataset (for instance, imputing 50 datasets if 50% of the data are missing). However, there is no fixed rule that will suit all research circumstances and the number of datasets could depend on factors such as the number of variables in the original dataset, the sample size, and the proportion of missing

[7] $$\sqrt{\frac{1}{M}\sum_k S_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k \left(b_k - \bar{b}\right)^2}\,, \qquad (4.5)$$

where:

b_k is the estimated regression coefficient in sample k of the M samples,

S_k the estimated standard error of the regression coefficient,

$\bar{b}$ the mean of b_k .

data (Shin, 2016). As a default, some statistical software (e.g., the TAM R package—Robitzsch, Kiefer, & Wu; 2019) use 10 imputed datasets.[8]

Secondly, when carrying out multiple imputation, it is important to include auxiliary variables into the imputation model (Shin, 2016). Auxiliary variables are variables like gender, school type, race or ethnicity—included in the imputation model to provide more information about the missing data—though not used in carrying out the IRT analysis (Shin, 2016). These variables exhibit a high bivariate correlation with the underlying missing data (Enders, 2010); and thus, increasing the number of these variables could increase chances of the MAR condition being satisfied in the dataset under investigation (Hardt, Herke, & Leonart, 2012; Schafer, 2003).

Even though multiple imputation addresses problems with conventional treatment methods (e.g. with respect to wastefulness, computational problems, biased [co]variances, and biased *p* values and confidence intervals), the method still suffers some deficiencies (Gingel, Linting, Rippe & van der Voort, 2019). For instance, the data must be missing at random, with the model used to generate the imputed values being "correct" in some sense (Allison, 2000). Importantly, the model used for generating the imputed datasets should match with the model used in carrying out the analysis (Allison, 2009; Rubin, 1987, 1996). That notwithstanding, multiple imputation remains one of the most popular missing data treatment methods amongst methodologists and researchers.

[8] For a more detailed discussion on the question about how many complete datasets to impute, see Graham *et al.* (2007).

Chapter 5 Study I—Effect of test length, sample size and population subgroups on measurement.

This chapter describes the first (of two large studies) carried out in this doctoral dissertation. It will begin with a background discussion on how the three factors (a) test length (b) sample size, and (c) population subgroups within the population, relate to the efficiency or precision with which population or item parameters are recovered. This will be immediately followed by a detailed description of the aim, methodology and description of the empirical and simulation study that was carried out.

5.1 Background

5.1.1 Test length and measurement precision or efficiency

Valid and reliable measures are crucial in the field of Psychology, as well as, in the study of abilities, aptitudes, and attitudes (Zanon, Hutz, Yoo & Hambleton, 2016). Further, it was long recognized that, all other things being equal, lengthening a test will increase its predictive validity due to the increased reliability of the test scores (Bell & Lumsden, 1980). This argument was derived from implications of the Spearman-Brown prophesy formula (Spearman, 1910), which defines the reliability of a test T, constructed by adding several n items $X_1,..,X_n$ as (Wolf, 2006, p. 24):

$$\mathrm{Rel}_T = \frac{n^2}{n(n-1) + \sum_{i=1}^{n} \frac{1}{\mathrm{Rel}(X_i)}}, \qquad (5.1)$$

However, the accuracy of the above formula depends on key assumptions like presupposing a specific "universe of content" from which random samples of items are drawn (Burisch, 1997). Further, if in the process of shortening a test, the deleted items are in every respect parallel to the remaining test items, the predicted validity of the shortened test, r_k, becomes accurately represented by the formula (Burisch, 1997, p.304):

$$r_k = \frac{r_0\sqrt{k}}{\sqrt{1+(k-1)r_{tt}}} \qquad (5.2)$$

with r_k representing the validity coefficient of the shortened test; r_0 the validity coefficient of the original test; *k*, the ratio of the new test length to the old test length; and r_{tt} the reliability coefficient of the original test.

Importantly, empirical studies confirm a positive association between test length and reliability (Wolf, 2006). For instance, Crotts, Zenisk, Sireci and Li (2013) evaluated the degree to which shortening tests in a multi-stage adaptive test impacted on measurement precision. They compared the test reliability from the original and reduced tests using several approaches and found that levels of measurement precision became better with reduced-length tests. On the other hand, test length and test quality are often found to be positively related when considering tests that are reference-based (Kruyen, Emons, & Sijtsma, 2012; Wilcox, 1980; Wolf, 2006). In such cases, the main concern is the agreement between decisions taken using outcomes of tests considered to be parallel (i.e., these parallel tests should classify subjects into categories in an identical manner).

Particularly, in an IRT setting, short tests can be more reliable than longer tests (Embretson, 1996). This occurs when the items of the short test are specially selected to be optimally appropriate to the ability or trait level of respective test takers (this is the case in targeted or adaptive testing). The reason for this is that in such targeted testing, the test items provide optimal information for the estimation of IRT model parameters, thus resulting in smaller standard errors of measurement for these tests (He & Wheadon, 2013). Conversely, if two tests—a long and a short test—have fixed item content (i.e., the same items are presented to

every test taker), the long test will produce lower standard errors of measurement than the shorter test (Embretson, 1996).

Yousfi (2005) used derivations from theory to explicitly prove that a positive relationship can only exist between test length and validity/reliability provided stringent conditions are met (e.g., test items being Rasch-homogenous or parallel). When these conditions are not satisfied, adding more items might not necessarily improve test validity or reliability. Thus, a straightforward monotone association between test length and test reliability (or validity) cannot be assumed. Two common error sources in the literature and statistical textbooks include researchers ignoring assumptions of the Spearman-Brown prophesy formula; and, failing to distinguish reliability/validity of an overall test and the reliability/validity estimates (Wolf, 2006). These errors lead to wrong conclusions about a test's statistical characteristics, and false associations between test length and test reliability (or validity).

Further, in real life testing scenarios, the assumption of parallel tests often gets violated. This is because, as more items become added to a test, there is a tendency for the response behaviour of test takers to change (e.g., test participants could become more bored or less motivated to complete a test, as more test items are added).

5.1.2 Sample size and measurement efficiency

The development of tests based on IRT and the analysis of data from such tests, rely heavily on the accuracy and stability of IRT model parameters. Also, when calibrating items from test data using IRT software, model parameter estimates and their associated standard errors (i.e., the measurement error associated with these estimated parameters) need to be provided. Since these IRT models are probabilistic in nature, an implication is that sample size will become a crucial factor impacting on how stable or accurate these estimated model parameters are (He & Wheadon, 2013).

Hambleton, Jones and Rogers (1993) carried out a simulation study using the 2-PL model and found that test information distributions became significantly different from their true

distributions as smaller sample sizes were used. Chuah, Drasgow and Leucht (2006) found similar results though using the 3-PL model and samples from the Computer Adaptive Sequential Test (CAST). Conversely, Wang and Chen (2005) used WINSTEPS (Linacre, 2006), to examine how sample size impacted fit statistics and standard errors of estimate when using the Rasch model (Rasch, 1960) and the Rating Scale model (Andrich, 1979). They found that the standard deviations of infit and outfit mean square errors (MNSQs) for the overall item difficulties became smaller in large samples. Similarly, DeMars (2003) examined how sample size, the total number of item parameters, and the number of parameters per item, influenced recovery of relevant parameters for polytomous items in a Nominal Response Model. Their results showed that sample size and the number of parameters per item accounted for a substantial amount of variance in RMSE, whereas the total number of item parameters did not.

The above studies used pure simulations in order to investigate effects of sample size on IRT parameter estimation. A limitation here is, assumptions underlying IRT models are perfectly satisfied, which rarely is the case when using test data from real life scenarios. However, a few studies examined this problem using operational test data. For instance, Swaminathan, Hambleton, Sireci, Xing and Rizavi (2003) employed a large dataset produced from the Law School Admissions Council test to examine the association between sample size and the specification of prior information on the accuracy of item parameter estimates. They found that when using small samples, incorporating ratings (provided by subject specialists and test developers regarding item difficulty in the form of a prior distribution) produced more accurate parameter estimates. Likewise, Stone and Yumoto (2004) used thirteen subsamples drawn randomly from the normative database of the latest edition of Knox's Cube Test Revised (KCT-R) to derive and compare estimates for the Rasch, 2-PL and 3-PL models. They found that as might have been expected, sample size influenced these estimates, with the Rasch parameter estimates from larger samples consistently having better goodness of fit indices than those from smaller samples.

Given that the above studies all use complete datasets, it might be interesting to investigate the effect of sample size on incomplete data sets—as is typically the case in most large scale

educational assessments, where the resulting incomplete datasets were treated with multiple matrix booklet designs. Moreover, a crucial objective in most large-scale assessments is precise recovery of population parameters of interest. Therefore, a desirable research objective could be to investigate how sample size is related to parameter recovery precision when using such multiple matrix booklet designs.

5.1.3 Measurement efficiency in policy relevant population subgroups

When carrying out large-scale educational assessments, another important objective is to investigate performance disparities between policy relevant sub-groups within populations (Schleicher, Zimmer, Evans, & Clements, 2009). Salient groups of persons could be classified in terms of relevant educational, geo-political or demographic variables like gender, ethnicity or socio-economic status. In the Unites States for instance, The Every Student Succeeds Act (ESSA) of 2015 (Public Law 114-15) demands state-wide accountability, wherein educational outcomes of students from major ethnic and racial groups, economically disadvantaged students, English learners, and students having disabilities are reported (Seastrom, 2017).

In several LSAs (for instance, PISA and NAEP), students' academic achievement distributions are estimated for policy relevant subpopulations or subgroups. This provides mean achievement scores and percentages of examinees lying above set cut scores within these in subpopulations or groups. This is done using sparse multiple matrix designs, which reduce respondent burden and test time, while ensuring broader content coverage. The resulting sparse data introduce large measurement errors for estimating individual ability estimates; thus, requiring special analysis procedures for estimating aggregate subgroup statistics (von Davier, 2003).

To tackle this methodology challenge, direct estimation is used, where subgroup estimates are obtained without generating individual achievement scores. This typically involves using "conditioning models" which incorporate responses given by students to cognitive items with their responses on background variables (Cresswell, Schwantner, & Waters, 2015). Thus, using additional student information from background variables, more accurate estimates of

subgroup characteristics are obtained—unlike when only responses to cognitive items are used (Yamamoto, Khorramdel, & von Davier, 2016).

Since 1984, the Educational Testing Service has been using a direct estimation technique for estimating relevant subgroup statistics. They use hierarchical IRT models to integrate student achievement data and background information, with major features of the procedure including (von Davier, 2003, p.1):

1. A population model with an assumption that students' proficiencies are normally distributed conditional on several background variables. Hence, marginal distributions (in general and for specific relevant subgroups) are a mixture of normals.

2. Generating a posterior latent trait distribution of proficiency for every subject in the sample using: an estimate from (1); the cognitive item responses; a set of IRT parameters estimated separately and treated as known and fixed; subjects' group membership; and other covariates. The combination of these individual posterior distributions gives an estimate of the true subgroup distributions.

3. Integrating over the posterior distributions of subjects and model parameters of the population—defined later in (1); to derive estimates of means, percentages over cut off achievement points, etc.

4. Using normal approximations for posterior distributions of subjects; and multiple imputation (using plausible values) to compute the integration in (3). The imputations are used in combination with conditioning models generated from cognitive response data and background information. These imputations are utilized to make it easier to perform the integration in (3) and to supply data which secondary data analysts can use with standard tools.

On the other hand, Cohen and Jiang (1999) put forward a different procedure for estimating subgroup characteristics using direct estimation. They argue that without using background variables, this approach yields consistent estimates for subgroup characteristics; with the main features involving (von Davier, 2003, p.2):

1. Population model generation with a marginal normality assumption (that is, aligning ability distributions from all subgroups will produce a joint normal distribution).
2. Categorical grouping variables have a measurement model with a continuous underlying variable and joint normal distribution with proficiency.
3. IRT model parameters are known and fixed.
4. Item responses are only used with a single grouping variable (used for reporting). No other covariates or their interactions are employed in the population model.
5. Direct calculation procedure which skips the generation of individual posterior distributions and plausible values.

Both methods described above are "direct estimation" methods as they do not use individual test scores to compute subgroup statistics. The method used by the ETS (described above) utilizes a more general model—with grouping variables, no assumptions concerning marginal proficiency distributions, and extra background information. The approach of Cohen and Jiang (1999) conversely applies a marginal normality assumption, while ignoring all background information except the single grouping variable (von Davier, 2003).

Using an approach like the ETS approach above, Gonzalez and Rutkowski (2010) carried out a simulation study using sparse multi-matrix designs and students' EAP scores, to investigate how item and population parameters were recovered in subgroup populations. They found that as the multi-matrix designs got sparser, the variance in the estimated posterior means for examinee proficiencies decreased, causing group differences to diminish. Thus, this resulted in real group differences being underestimated with this degree of underestimation increasing noticeably, as fewer items were administered per examinee (i.e., as matrix sparseness increased).

However, even though accurate recovery of performance gaps between policy-relevant population subgroups remains a key objective in large-scale assessments, few studies have critically examined this question; especially, how factors such as test length and sample size, relate to the precision with which subgroup or subpopulation parameters are recovered.

5.2 Research objectives and research questions

The results of large-scale student assessment studies inform evidence-based educational policies and significantly contribute to the empirical knowledge base of educational research. Multi-matrix booklet designs in conjunction with IRT analyses represent the state-of-the-art methodology of large-scale assessment studies. Thus, efficient and unbiased parameter recovery of population parameters or parameters concerning policy-relevant performance gaps related to gender, school type or immigration background are key evaluation criteria of this methodology. However, the knowledge base on multi-matrix designs and parameter recovery is still limited (Rutkowski, Gonzalez, von Davier & Zhou, 2014). Some simulation studies on multi-matrix booklet designs (e.g., Gressard & Loyd, 1991; Gonzalez & Rutkowski, 2010)[9] showed that parameter recovery of population and subgroup parameters are influenced by factors such as test length and the number of participating students (i.e., sample size). Also, most large-scale assessment studies apply the balanced incomplete block booklet design.

On the other hand, although the excellent studies mentioned above significantly contributed to the knowledge base on parameter recovery of multi-matrix sampling designs, there are still important research gaps. Notably, these studies were carried out using completely simulated data. Such data has well spelt out characteristics and fits to the underlying statistical models in predefined ways (e.g., perfect fit). However, this is not true for real empirical data from operational testing situations, where IRT models only provide an approximation to the empirical data and hence different results might be expected. Thus, it is not clear whether the results of simulation studies can be generalized to real empirical data. Further, it is also interesting to systematically investigate the extent to which parameter recovery precision is lost as booklet designs get sparser—as well as, considering test length and sample size.

This study therefore used (a) real assessment data and (b) simulated data to critically examine how the factors test length and sample size influence parameter recovery when using different balanced incomplete block booklet designs. These booklet designs differ in their

[9] These studies were detailly described in chapter one—See Section 1.2 of this dissertation.

levels of matrix sparseness (i.e., the amount of missing data they contain). To this end, real and simulated data was rigorously examined where all or some of the factors—test length, sample size and matrix sparseness—are manipulated experimentally. In particular, the following research questions were tackled: (1) How precisely can item and person parameters be recovered at the global population level when using these different sparse booklet designs[10]? (2) How do test length or sample size influence the precision (or efficiency) with which person and item parameters are recovered when using these booklet designs[11]? (3) How precisely can parameters related to performance differences of policy-relevant subgroups (e.g., gender, migration background and school type) be recovered in these booklet designs?

5.3 Data and procedure

To tackle the above research questions, this study drew *real data*, as well as *simulated data.* (The *simulated data* was generated to have properties of the real dataset).

Real data was obtained from the 2015 VERA-8 Mathematics Assessment for the German federal states of Berlin and Brandenburg. VERA-8 is a yearly assessment which assesses the mathematics achievement of 8th Graders and run by the Institute for School Quality for Berlin and Brandenburg (ISQ). Content specialists construct the test items with strict adherence to German national educational standards and evaluate psychometric properties of items by pre-testing them in large field studies. This assessment was chosen because it gives crucial information to teachers, school principals and education policy makers, and since the data are not simulated, but rather typical of large-scale assessments with respect to their fit to underlying IRT models (in the case of this study, the unidimensional 1-PL IRT model, aka Rasch Model). This *real dataset* contains responses of 13,076 students from the non-academic school track on a test of 48 dichotomously scored items. The student population

[10] In this research question, the real dataset of 42 items and 10,000 students, randomly selected from the VERA-8 dataset was used.
[11] In this research question, all test lengths (42, 84, 126 items) and sample sizes (300, 500, 1000, 3000, 4500, 6000, and 10,000 students) were considered.

comprises 45% female; 74% German, 9% Turkish, 17% other nationalities; students attend either the integrated secondary school (71%), or the integrated comprehensive school ("Gesamtschule", 19%). These two school types differ in their pedagogical orientation.

Simulated data were generated using the R package *irtoys* (Partchev, 2017). This was done by using (a) item difficulties from the real dataset and (b) the mean and standard deviation of student abilities from the real dataset, assuming a normal distribution, to generate student response data. Response data were simulated for test lengths of 84 and 126 items. These numbers were chosen because they are plausible item numbers that could be used in operational test situations; and since the numbers are all multiples of 7, they can be used in creating BIB7 booklet designs. Also, since BIB7 booklet designs can only be created when the test length is a multiple of 7, subjects' responses to 42 items were randomly selected from the original real dataset. (See Table 5.1 for a structural overview of the experimental conditions and variables used in this study).

To learn how sample size affects parameter recovery studies (a) real, (b) simulated data was studied for various sample sizes: 300, 500, 1,000, 3,000, 4,500, 6,000 and 10,000 students. These numbers chosen because they cover a wide spectrum of the possible number of students that can partake in an operational test (e.g., PISA requires a minimum sample size of 4,500 students from each participating country).

To this end, random selection of students from the original VERA dataset of 13,076 students was performed for the real and partly simulated data, while person and item characteristics of the VERA dataset were used in creating simulated datasets.

Further to tackle the question on the recovery of group differences between policy relevant sub-groups, data like in the VERA dataset was simulated for a population of two groups—Group 1 and Group 2. This was done using the item difficulties of the VERA-8 dataset and assuming the variance of ability in both groups was 1. These two groups represent policy relevant sub-groups (for instance, consisting of students with different genders, migration backgrounds, or socio-economic backgrounds).

Table 5.1. *Overview of the study design*

Data to be studied	Experimental Conditions				Major Dependent Variables
	Booklet Design	Sample Size	Total Number of Items	Difference in mean person ability between population sub-groups	
Real data	0,1,2,3	300, 500, 1,000, 3,000, 4,500, 6,000, 10,000	42	--	Bias & RMSE for item and person parameters at the population level
Simulated Data	0,1,2,3	300, 500, 1,000, 3,000, 4,500, 6,000, 10,000	84, 126	--	Bias & RMSE for item and person parameters at the population level
Simulated Data	0,1,2,3	300, 500, 1,000 3,000, 4,500 6,000 10,000	42	0[a], 0.4[b], 0.8, 1.2, 1.6 2	Bias & RMSE for recovered difference in person parameters between population subgroups

Note. The booklet design (see Section 5.4 below) determines the number of items per student, i.e., the matrix sparseness; with booklet design 0/1/2/3 implied that a student worked on 42/18/12/6 items, respectively. All experimental conditions were fully crossed. The simulated data was created using item and person characteristics from the original VERA dataset used in the study. The person parameters investigated were the mean person ability and the variance of person abilities, while the item parameter examined were the mean item difficulty.

[a] Data was simulated for two population subgroups assumed to be policy relevant (Group1 and Group2). In this simulation condition, Group1 and Group2 both have mean person abilities of 0 (Hence, no group difference in their mean person abilities) with standard deviation of 1.

[b] In this condition, Group1 has a mean person ability of -0.2, while Group2 has a mean person ability of 0.2, giving a group difference in mean person ability of 0.4). Similarly, to get group differences (for the mean person ability) of 0.8, 1.2, 1.6 and 2; the mean person abilities for Group1 and Group2 was simulated to be (-0.4 and 0.4), (-0.8 and 0.8), and, (-1 and 1) respectively. Also, the group differences are given on the logit scale, with variance of person ability distribution being 1 in every group (Each group consists of 5000 students).

All experimental conditions were fully crossed. Doing so, yields 4 booklets (described in detail later below) x 7 (sample size) = 28 experimental conditions for first research question

with real data; 4 (booklets) x 7 (sample sizes) x 2 (total number of items) = 56 conditions for the second research question with simulated data; and, 4 (booklets) x 1 (sample size) x 6 (levels of group differences in mean person ability) = 24 conditions for the third research question with simulated data.

To create the real datasets required to answer the first research question, simple random sampling of 42 items and the appropriate number of students were selected from the original VERA dataset. So, depending on the experimental condition either 300, 500, 1000, 3000, 4500, 6000 or 10000 students were randomly sampled from the real VERA-8 dataset.

To create the simulated datasets required to answer the second research question (i.e., for test lengths of 84 and 126 items), the mean and standard deviations of person ability and item difficulties from the VERA-8 dataset were used to generate response data for the appropriate number of students (i.e., for either 300, 500, 1000, 3000, 4500, 6000, or 10000 students). Similarly, to create simulated data to answer the third research question (where the test length is 42 items), item difficulties from 42 randomly selected VERA-8 items were used to generate response data for an appropriate sample size where the mean difference in person ability between Group 1 and Group 2 was simulated to be either 0, 0.4, 0.8, 1.2, or 2 (See note in Table 5.1 above for the procedure in which these mean differences in person ability were simulated). The choice of sample sizes, test lengths and group differences in mean person ability for the various experimental conditions were chosen because they are plausible in operational educational assessments.

The items used fit well to the Rasch model, with Table 5.2 below showing fit statistics for 42 randomly selected VERA-8 items and aggregated across 1000 replications. These values show a good fit to the Rasch model as the infit and outfit values are always close to 1.

Thus, in this study a total number of 28 + 56 + 168 = *252 experimental conditions* were investigated. Further to get stable estimates each of these conditions was replicated 1000 times producing 1000 datasets from which the parameter of interest was calculated.

Table 5.2. Item fit statistics for VERA-8 items

	Item Name	Outfit	Outfit_t	Infit	Infit_t
1	M1620701	1.12	7.49	1.02	1.03
2	M1620702	1.02	1.86	0.99	-0.61
3	M1642601	0.99	-0.69	0.98	-1.46
4	M1642602	1.08	10.59	1.06	7.54
5	M1640101	1.02	3.25	1.01	1.75
6	M1511401	0.99	-0.32	0.96	-2.34
7	M1511402	0.93	-17.88	0.95	-12.45
8	M1641201	1.03	3.69	1.01	1.21
9	M1621802	1.19	10.48	1.03	1.56
10	M1632201	1.25	29.43	1.17	20.73
11	M1641001	1.04	4.22	0.99	-1.49
12	M1632501	0.96	-17.32	0.94	-12.86
13	M1632502	1.05	3.39	1.03	2.34
14	M5645101	1.16	8.35	1.07	3.49
15	M5645102	1.27	24.34	1.14	12.50
16	M5642201	1.09	3.84	1.00	-0.01
17	M5640601	0.98	-6.48	0.94	-3.03
18	M5640602	1.14	17.66	1.09	11.18
19	M5640603	0.92	-25.66	0.96	-19.39
20	M5640501	1.20	13.13	1.07	5.02
21	M5640503	0.93	-9.11	0.96	-5.83
22	M5642001	1.10	12.36	1.07	8.40
23	M5640401	0.92	-16.81	0.91	-11.91
24	M5640403	0.96	-15.00	0.97	-10.30
25	M4631601	1.03	1.51	0.98	-1.35
26	M4641901	0.91	-18.42	0.90	-13.45
27	M4641903	1.01	0.72	1.01	1.06
28	M4641001	0.93	-9.96	0.96	-5.69
29	M4641002	1.14	17.69	1.08	10.76
30	M4641101	0.99	-0.62	1.00	0.03
31	M4641102	0.91	-14.60	0.92	-8.26
32	M4641201	1.10	7.07	1.02	1.39
33	M4645001	0.89	-11.35	0.93	-7.00
34	M4610402	0.98	-1.53	1.01	0.87
35	M4642101	0.91	-16.26	0.93	-8.40
36	M4642102	1.17	20.54	1.14	17.29
37	M2645301	1.11	7.63	1.04	3.17
38	M2500301	1.05	3.09	0.99	-0.46
39	M2645001	0.97	-1.37	0.95	-2.81
40	M2641401	0.99	-1.70	0.99	-1.70
41	M2641901	0.98	-2.18	0.97	-3.09
42	M2641902	0.97	-3.59	0.98	-2.31

Note. Results based on 42 randomly selected test items.

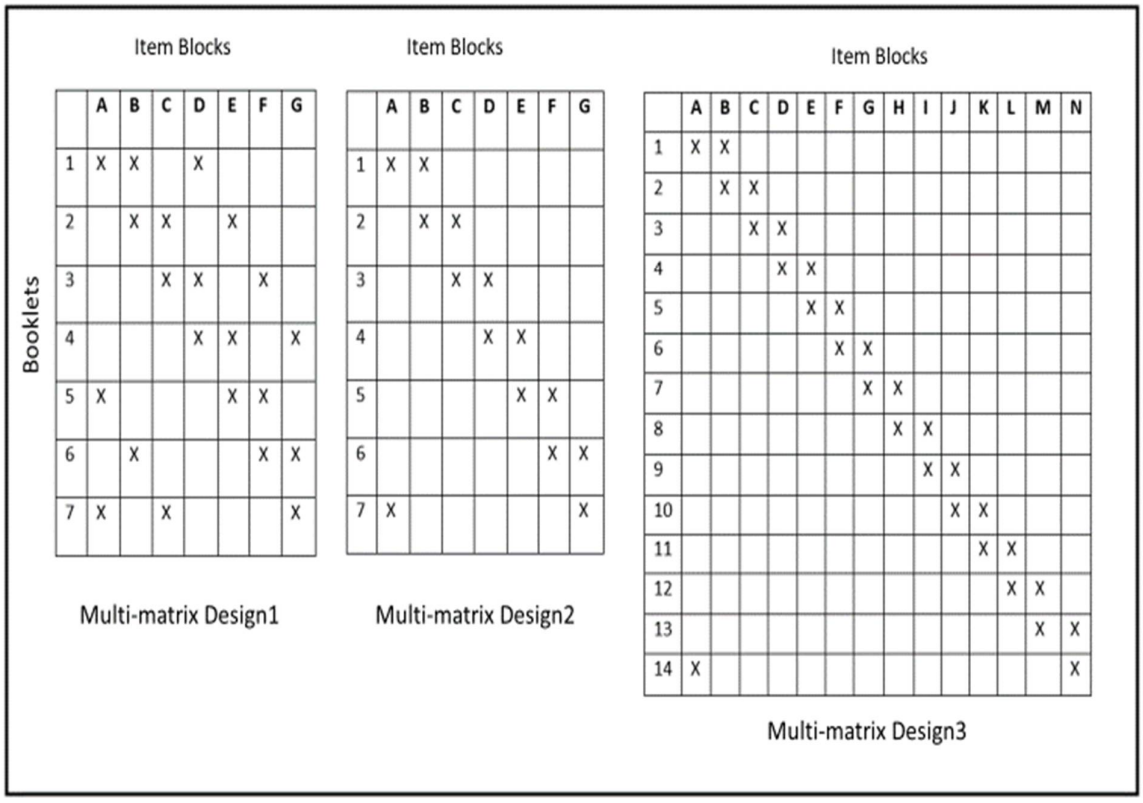
Drawing on the methodological approach applied by von Davier *et al* (2009) as well as Gonzalez and Rutkowski (2010), various booklet designs were constructed using sparseness techniques. Hence, the key parameter in which these booklets differ is the number of items

answered by each individual student. Since the BIB-7 block design can only be created on a test with a test length which is a multiple of 7, data for 42 items from the original VERA-8 data were randomly selected. The following booklet designs were analysed with real data from a total number of 42 items, as well as with partly and fully simulated data with 84 items, 126 items, and 168 items.

In the following the booklet designs with a total number of 42 items is described. The booklet designs with a larger number of items were created analogously.

- *Booklet Design0*: The 42 test items were presented to all examinees. Given that all students worked on all 42 items, this complete booklet design served as the gold standard for evaluating parameter recovery of the other incomplete designs.
- *Booklet Design1*: Items were randomly distributed into one of seven blocks, labelled A, B, C, D, E, F, and G. Based on the BIB7 design, every examinee is administered three blocks containing 18 items from the assessment pool. Hence the blocks were organized as shown in Figure 5.1 below with the resulting booklets being (ABD), (BCE), (CDF), (DEG), (EFA), (FGB), and (GAC). According to Frey, Hartig & Rupp (2009) this booklet design is characterized by the following parameters: t = 7, b = 7, r = 3, k = 3, λ = 1. This booklet design contains 57% missing data (i.e. 24/42 *100, for the 42-itemed test, since 24 test items are not administered to the test taker).
- *Booklet Design2*: Every examinee responds to two blocks containing 12 items from the assessment pool; the blocks used were like those in Booklet Design1 above. The blocks were arranged into seven pairs thus: (AB), (BC), (CD), (DE), (EF), (FG), and (GA). Hence, the design parameters were t = 7, b = 7, r = 2, k = 2, λ = 1. This booklet design contains 71% missing data (i.e., 30/42 * 100, for the 42-itemed test, as 30 items are not administered to the test taker).
- *Booklet Design3*: Items were randomly distributed into one of 14 blocks (i.e., blocks A through block N). Every examinee responds to two of these blocks containing six items from the assessment pool. These blocks were arranged into 14 pairs thus: (AB), (BC), (CD), (DE), (EF), (FG), (GH), (HI), (IJ), (JK), (KL), (LM), (MN), and (NA). The design parameters for this last case are therefore: t = 14, b = 14, r = 2, k = 2, and λ = 1. This booklet design contains 86% missing data (i.e., 36/42 * 100, for the 42-item test).

Fig. 5.1. *An overview of the various booklet designs used in the study*



Note. Booklet design0 is the complete data design, where all students are administered all test items. The design thus serves as the gold standard for comparing the other booklet designs.

5.4 Method of data analyses

Given that no standard software or packages are available, test item selection algorithms were newly programmed to generate the multi-matrix booklet designs used in the study. The programming was done and run in the R environment for statistical computing (R Development Core Team, 2017). First, items and persons in the original VERA-8 dataset were randomly selected to create new datasets. Next, test items were ordered in increasing difficulty; after which three equally sized items groups (easy, average, and difficult items) were sequentially created. To create item blocks for the various booklet designs, an equal number of items were randomly selected from each of the item groups (i.e., easy, average and

difficult items). This was to ensure homogeneity of item difficulty in the different item blocks. Conversely, persons were equally divided (sequentially) into 7 blocks (for booklet Design1 and Design2) or 14 blocks (for booklet Deign3). Details of the algorithms used in creating these booklet designs are shown in Appendix B.1.

Person and item parameters were scaled using the TAM R package (Robitzsch, Kiefer, & Wu, 2017) with the Mixed Coefficients Multinomial Logit Model (MCMLM; Adams, Wilson, & Wang, 1997; Adams & Wu, 2007). The MCMLM is a generalized multi-dimensional Rasch model which allows estimation of multi-dimensional distributions conditional on background variables. Introducing background variables with the use of a latent regression permits estimation of not only first moments (i.e., means) and second moments (i.e., standard deviations) of overarching multidimensional distributions, but also moments of subgroup-specific distributions nested in the overarching distribution (Frey & Bernhardt, 2012). This allows means and variances of students from specific population subgroups (e.g., based on gender, school type, or migration background) to be easily calculated.

Multiple Imputation (Rubin, 1987) was used to estimate person parameters. The technique involves substituting every missing data point with a set of *m* plausible values to create *m* complete data sets. Standard statistical software is then used to analyse these complete datasets and combining the results to get parameter estimates (Sinharay, Stern & Russell, 2001). The TAM R package (Robitzsch, Kiefer, & Wu, 2017) was used in carrying out the multiple imputation, with 10 plausible values (which is the default in TAM) and as applied in many large-scale assessments (Frey & Bernhardt, 2012). Importantly, multiple imputation with plausible values was used because it remains state-of-the-art methodology in analysing large-scale assessments of student performance (Braun & von Davier, 2017).

On the other hand, although many indices exist for evaluating statistical properties of booklet designs with respect to simulation studies, two most popular indices are the bias and RMSE (e.g., Toland, 2008; Svetina *et al*, 2013; Custer, 2015; and, Hecht *et al*, 2015a). Hence, these two indices were used to evaluate the efficiency with which person and item parameters are recovered in booklet designs. Bias describes the mean difference between the true and

estimated parameter values (i.e., mean inaccuracy of parameter estimate); while, Root Mean Squared Error (RMSE) signifies the root of the average squared difference between estimated and corresponding true parameter value (Wolf, 2006). Thus, RMSE considers variability of parameter estimate, with smaller RMSE and bias values implying more accurate or better parameter recovery. Further, given N replications, RMSE and bias are computed as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2}\,, \qquad (5.1)$$

$$Bias = \frac{\sum_{i=1}^{N}(X_i - \hat{X}_i)}{N}\,, \qquad (5.2)$$

where X_i is the true parameter value, $\hat{X}_i$ the estimated value of the parameter, and N the number of replications to be carried out (In this study, N = 1000). With the real dataset, the true parameter values are obtained from Booklet0, which is the complete dataset; while this value is known when defining the parameter value for the simulated datasets.

Specifically, to calculate the RMSE and the bias for the mean of the person ability distribution, the mixed coefficient multinomial logit model was applied to each simulated dataset with 10 PVs. This resulted in 10 sets of PVs each having a mean. The average of these 10 means were taken to represent an estimate for the person parameter (i.e., the mean of the distribution of person ability) for that given dataset. This procedure was then replicated using 1000 different (simulated) datasets, resulting in 1000 estimates for this person parameter. These 1000 estimates of the recovered person parameter are now compared with the true parameter value got from the complete dataset (where no multi-matrix design was applied). This same procedure was followed to calculate the RMSE and bias for the mean item difficulty. However, to compute the RMSE and bias for the variance of the distribution of person ability, the procedure to calculate the variance of the distribution of variance of person ability using PVs as described in Section 3.3 of this dissertation was observed (i.e., the imputation variance was added to the variance of PVs for each set of 10 PVs for any given

dataset). Further, for interpretation purposes, a parameter was considered accurately recovered when RMSE ≤ 0.04. This approximately corresponds to the acceptable standard error benchmark in PISA, where the estimated parameters should lie within ± 5 PISA points from their true values and thus, typically falling within a magnitude of 2 standard errors of sampling (OECD, 2014, p.27).

Importantly, in addition, inferential statistics were reported following very closely, the recommendations of Harwell, Stone, Hsu, & Kirisci (1996) and Feinberg & Rubright (2016), for reporting results of simulation studies performed in Item Response Theory and Psychometrics. These simulation results were "summarised as ANOVA's to highlight the main effects" (See, Feinberg & Rubright, 2016, p. 44). Also, as recommended by Harwell et al. (1996, p.21), a non-linear, log transformation of the dependent variable (i.e., RMSE) was performed to increase the likelihood of the normality assumption being satisfied. Thus, since the item difficulties were normally distributed, it is expected that the log(RMSE) of parameters computed from these item parameter should be asymptotically normally distributed, with a mean and variance depending on the number of replications (Bartlett & Kendall, 1946; Harwell et al., 1996). In the same light, effect sizes of independent variables were computed using eta squared. As emphasized by Levine & Hullett (2002, p. 612), "Eta squared (η^2) is the most commonly reported estimate of effect size for the ANOVA". Particularly, η^2 is easy to interpret, as it represents the percentage of variance associated with each independent variable; and all sources of variation (with their individual errors) add up to 1.00.

5.5 Results and discussion

5.5.1 How efficiently are item and person parameters recovered at the global population level when using the different sparse matrix booklet designs?

The person parameters investigated were the mean and the variance of the distribution of person abilities, while the item parameter investigated was the mean item difficulty. Using the original VERA dataset (with complete data and no multi-matrix design applied), the mean of the distribution of person abilities was 0.00, while the mean item difficulty was -0.70. Similarly, the variance of the distribution of person abilities was 1.17, while the variance of the item difficulties was 1.19. Importantly, the mean and the variance of the distribution of person abilities were recovered accurately in all sparse booklet designs (RMSE $\leq$ 0.04), with the mean of the distribution of person abilities recovered more accurately than the variance of the distribution of person abilities. Further, the mean of the distribution of item difficulties was only accurately recovered in Booklet Design1 and Booklet Design2. However, there was no bias in the recovery of all person and item parameters investigated ($0.00 \leq$ Bias ≤ 0.01). Table 5.3 shows detailed values of the RMSE and bias for the recovery of person and item parameters across the various booklet designs.

To further illustrate how well person abilities were recovered, the distribution of person abilities in the original VERA dataset, was compared to the distribution of person abilities from the datasets to which the various sparse matrix booklet designs had been applied. Figure 5.2 shows this result for the recovered ability distributions using the first plausible values of ability (more detailed results, for the other plausible values, can be found in Appendix A.1). The results show that all designs recover the original distribution of person abilities very well (except around the centre of the ability distribution where the recovered distributions deviate slightly from the distribution in the original dataset, as the booklet designs became sparser).

For each experimental condition, 1000 datasets were generated using the method described in section 5.3, from which a pooled estimate of the parameter of interest (for instance, the mean of the distribution of person abilities) was computed. Figure 5.3 and Figure 5.4 show how these parameters were recovered across the 1000 datasets for the various sparse matrix

booklet designs. The results clearly show that as the number of items administered per student became fewer (i.e., the booklet design became sparser), the precision with which the parameters of interest were recovered reduced.

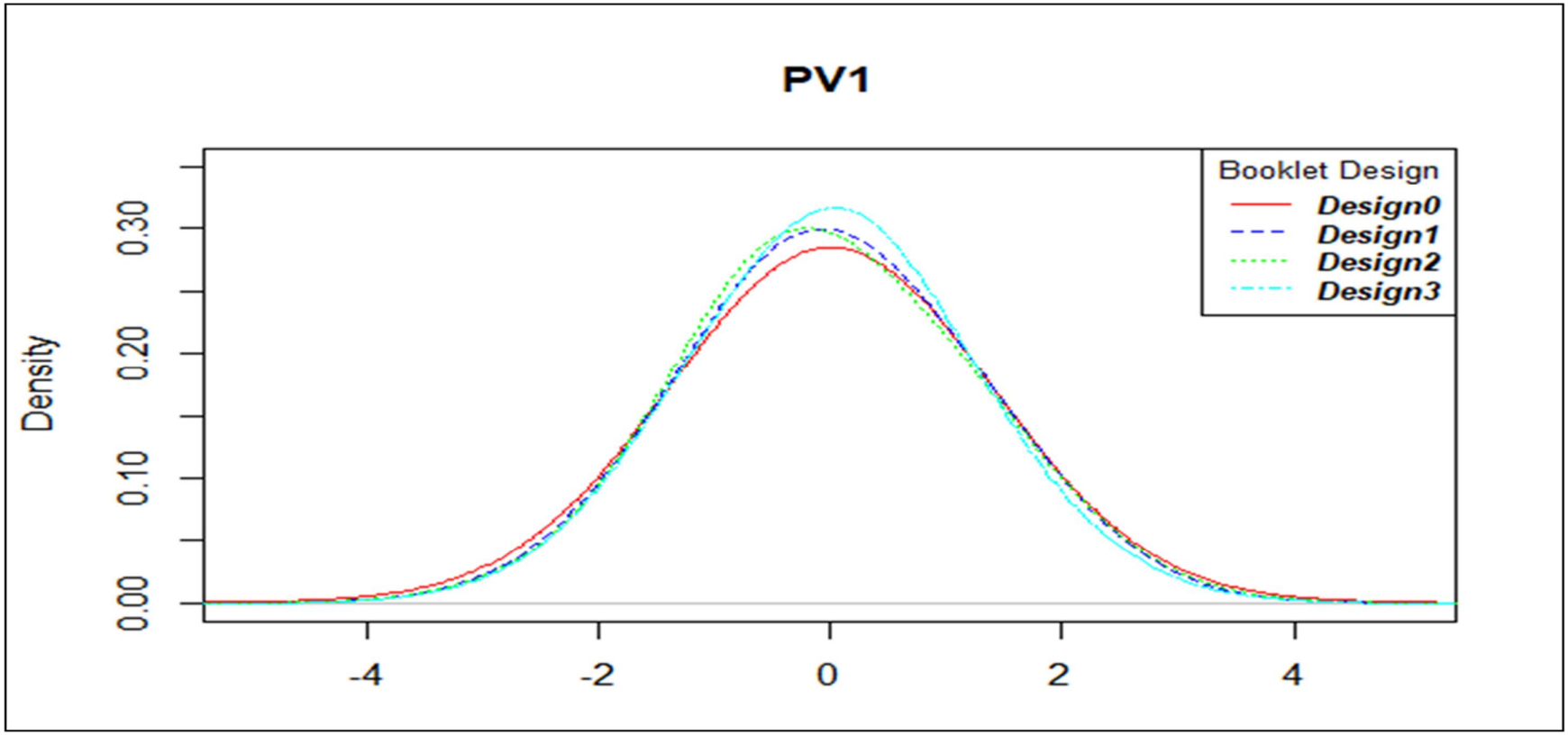
Additionally, to examine how individual item locations were recovered, an extra analysis was performed to calculate RMSE's and bias for 42 randomly selected items from the VERA-8 test. The results showed that RMSE's and bias for item location parameters increased as the booklet designs became sparser. However, no clear pattern was found in the recovery of the item locations based on their difficulty (i.e., after arranging the items in order of increasing item difficulty, no clear relationship was found between the difficulty of an item and the precision with which its difficulty was recovered). Details of the RMSE and bias values for these individual items (across 1000 replication conditions) are shown in Table 5.4. The results for the RMSE's of recovered item locations per item are summarized in Figure 5.5; while Figure 5.6 summarizes these details for the bias.

Table 5.3. *Parameter recovery efficiency for person and item parameters across booklet designs when using the VERA 2015 mathematics dataset*

	Complete data design	Booklet Design1	Booklet Design2	Booklet Design3
Mean Person Ability	-0.0038	-0.0065	-0.0073	-0.0079
RMSE mean person ability	-	0.0061	0.0071	0.0081
Bias mean person ability	-	0	0	0.0001
Variance Person Ability	1.3926	1.3901	1.3897	1.3881
RMSE variance person abilities	-	0.0229	0.0241	0.0317
Bias variance person abilities	-	-0.0013	-0.0027	-0.0032
Mean item difficulty	-0.7015	-0.707	-0.7128	-0.7213
RMSE mean item difficulty	-	0.0273	0.0409	0.1017
Bias mean item difficulty	-	-0.0026	-0.0039	-0.0065

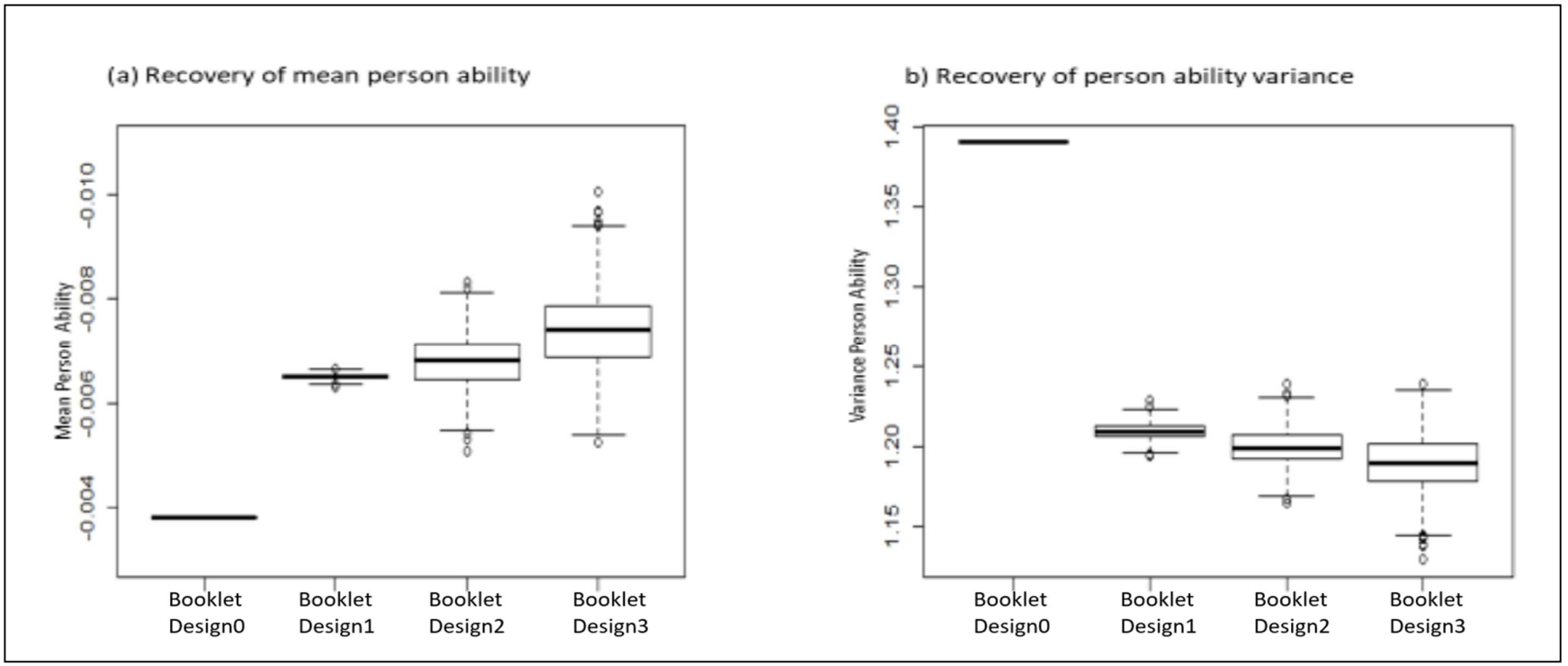
Note. In the complete design no matrix sampling was used. For Booklet Designs 1 through Booklet Design3, the results obtained were pooled from 1000 iterations. Progressively, the booklet designs become sparser moving from Booklet Design1 to Booklet Design3. (N = 10,000 students). For each simulation condition, the three parameters—mean person ability, variance of person abilities, mean item difficulty—were computed for each of the booklet designs from which the respective RMSE and bias were computed.

Figure 5.2. *Recovery of the distribution of person abilities*



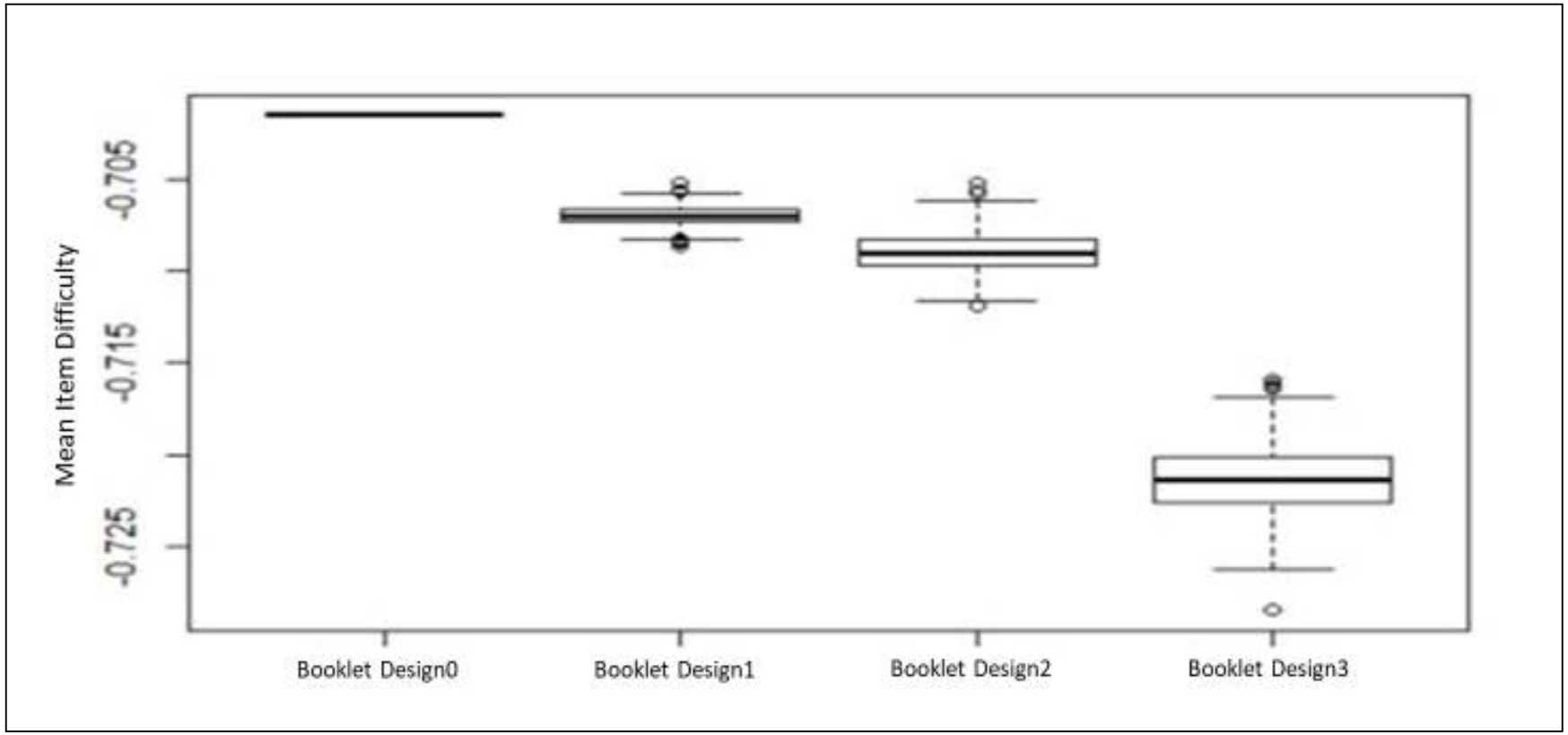
Note. The recovery of the distribution of person abilities is from the 2015 VERA8 Mathematics dataset for Berlin and Brandenburg. The figure shows recovery of this distribution using the first plausible value and for the various booklet designs. Design0 contains the complete dataset having no missing data, while the designs become sparser moving from Design1 to Design3. (N = 10, 000 students).

Figure 5.3. *Recovery of the mean and variance of the distribution of person abilities*



*Note***.** The results for the complete dataset (i.e., Design 0) are obtained from a single computation, while those for Booklet Designs1, Design2 and Design3 are obtained from 1000 iterations. (N = 10, 000 students).

Figure 5.4. Recovery of the mean item difficulty across various booklet designs



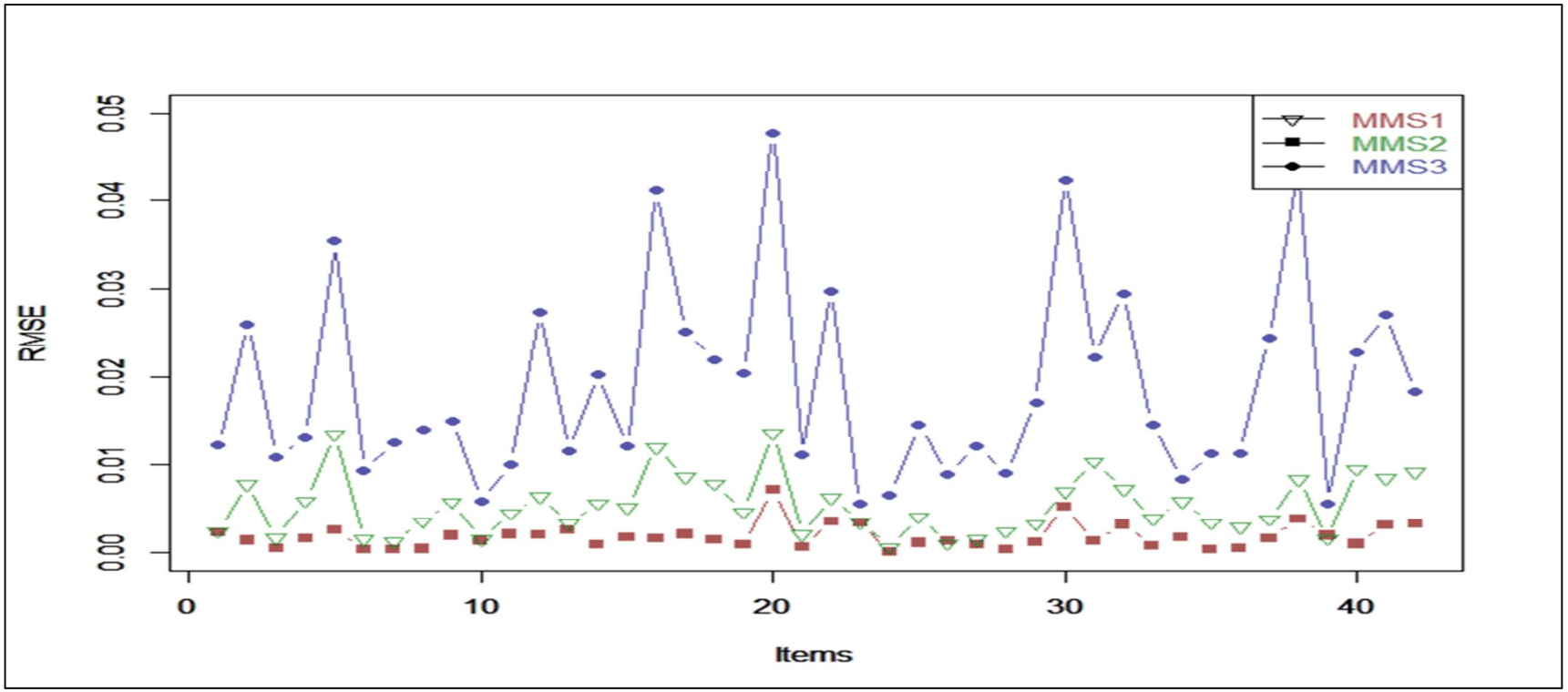
Note. The results for the complete dataset (Booklet Design0) are obtained from a single computation, while those for Booklet Designs1, Design2 and Design3 are obtained from 1000 iterations. (N = 10, 000 students).

Table 5.4. RMSE and Bias for recovery of item location parameter across items in the VERA-8 dataset

	RMSE			Bias		
Items	MMS1	MMS2	MMS3	MMS1	MMS2	MMS3
I1	0.0023	0.0025	0.0123	0.0084	0.0110	0.0392
I2	0.0014	0.0078	0.0260	-0.0080	-0.0220	0.0042
I3	0.0005	0.0017	0.0109	0.0028	0.0033	0.0077
I4	0.0016	0.0060	0.0131	0.0077	0.0053	0.0193
I5	0.0026	0.0134	0.0354	-0.0060	-0.0175	-0.0047
I6	0.0003	0.0015	0.0094	-0.0008	-0.0002	-0.0045
I7	0.0003	0.0013	0.0126	0.0003	-0.0027	-0.0081
I8	0.0005	0.0036	0.0139	0.0006	0.0099	0.0181
I9	0.0020	0.0058	0.0150	-0.0065	-0.0116	-0.0186
I10	0.0013	0.0015	0.0059	0.0024	0.0071	0.0138
I11	0.0021	0.0046	0.0101	0.0005	0.0081	0.0024
I12	0.0021	0.0064	0.0274	0.0057	0.0136	0.0259
I13	0.0026	0.0034	0.0116	0.0001	0.0097	0.0029
I14	0.0010	0.0056	0.0203	0.0044	-0.0060	0.0007
I15	0.0018	0.0052	0.0121	-0.0064	-0.0146	-0.0219
I16	0.0016	0.0120	0.0412	-0.0050	-0.0200	-0.0361
I17	0.0021	0.0087	0.0252	0.0026	-0.0030	-0.0020
I18	0.0014	0.0079	0.0220	-0.0019	-0.0097	-0.0136
I19	0.0010	0.0047	0.0204	-0.0047	-0.0072	-0.0018
I20	0.0072	0.0136	0.0478	0.0209	0.0264	0.0325
I21	0.0007	0.0021	0.0112	0.0034	-0.0023	0.0018
I22	0.0036	0.0063	0.0298	0.0070	-0.0022	0.0100
I23	0.0035	0.0035	0.0056	0.0193	0.0087	-0.0002
I24	0.0002	0.0007	0.0066	0.0012	-0.0029	0.0141
I25	0.0011	0.0041	0.0145	-0.0027	-0.0078	-0.0227
I26	0.0013	0.0011	0.0090	0.0120	0.0097	0.0053
I27	0.0009	0.0016	0.0122	0.0024	-0.0008	-0.0034
I28	0.0004	0.0025	0.0090	-0.0012	-0.0057	0.0054
I29	0.0013	0.0033	0.0171	0.0040	-0.0013	0.0090
I30	0.0053	0.0070	0.0423	0.0145	0.0000	-0.0036
I31	0.0013	0.0104	0.0222	-0.0004	-0.0072	-0.0103
I32	0.0032	0.0073	0.0295	0.0027	0.0063	0.0093
I33	0.0008	0.0039	0.0146	0.0076	0.0135	0.0160
I34	0.0018	0.0059	0.0084	0.0059	0.0060	-0.0045
I35	0.0004	0.0034	0.0113	0.0030	0.0006	0.0088
I36	0.0006	0.0030	0.0113	-0.0039	-0.0130	-0.0143
I37	0.0016	0.0039	0.0244	-0.0118	-0.0084	-0.0246
I38	0.0039	0.0084	0.0438	0.0046	-0.0058	-0.0147
I39	0.0020	0.0017	0.0057	-0.0118	-0.0019	-0.0170
I40	0.0010	0.0096	0.0227	-0.0092	-0.0293	-0.0425
I41	0.0032	0.0085	0.0271	0.0065	0.0037	0.0052
I42	0.0033	0.0092	0.0183	-0.0080	-0.0202	0.0053

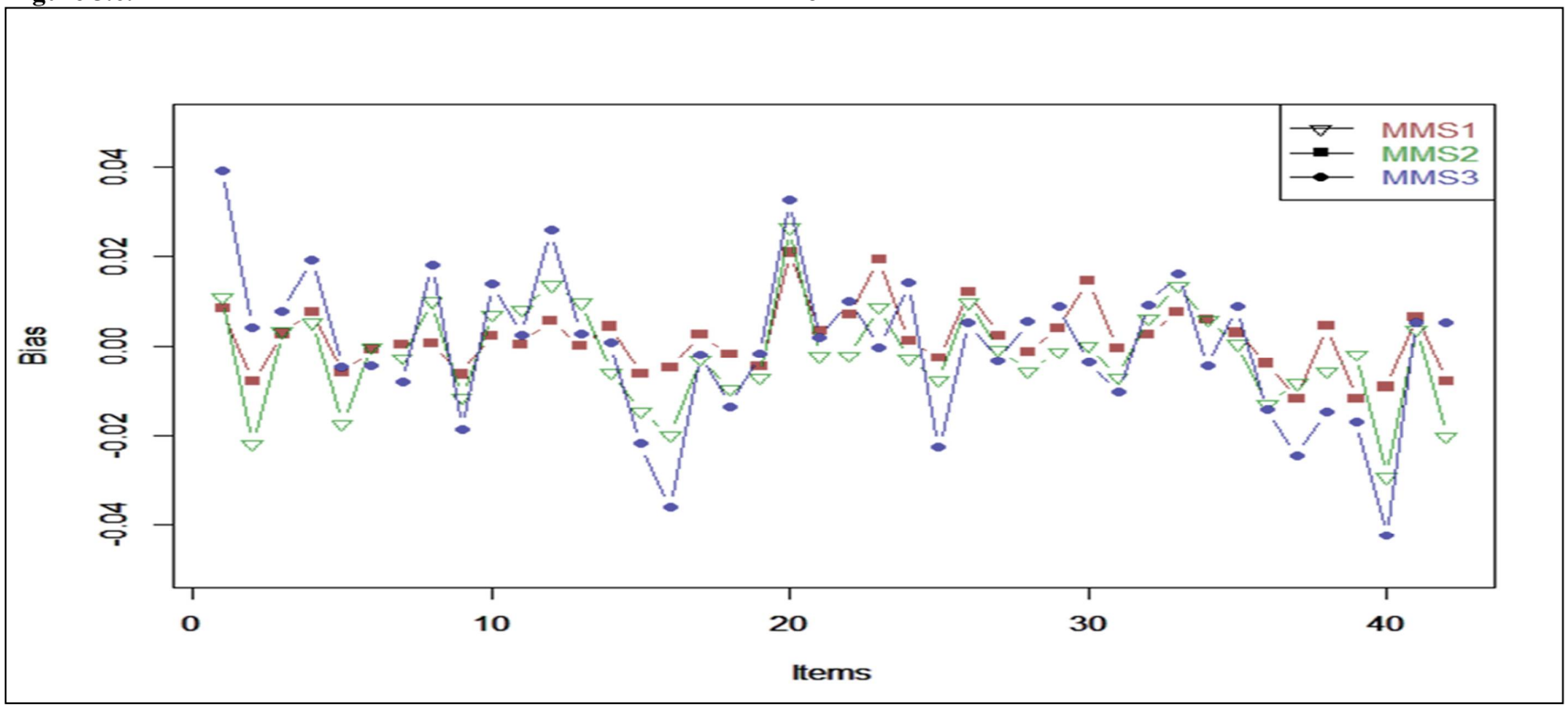
Note. MMS1/MMS2/MMS3 represent the multiple matrix booklet Design1/Design2/Design3 respectively. (N = 10, 000 students).

Figure 5.5. RMSE for recovered item locations at item level in the VERA-8 dataset



Note. The results are for 42 randomly selected items from the VERA-8 dataset and for 1000 replication conditions. Each point represents a single item and these items are arranged in increasing order of difficulty. MMS1/MMS2/MMS3 represent multiple matrix booklet Design1/Design2/Design3 respectively. (N = 10, 000 students).

Figure 5.6. Bias for recovered item locations at item level in the VERA-8 dataset



Note. The results are for 42 randomly selected items from the VERA-8 dataset and for 1000 replication conditions. Each point represents a single item and these items are arranged in increasing order of difficulty. MMS1/MMS2/MMS3 represent multiple matrix booklet Design1/Design2/Design3 respectively. (N = 10, 000 students).

5.5.2 How is test length and sample size related to the efficiency or precision with which person and item parameters are recovered in the various sparse matrix booklet designs?

The person parameters investigated were the mean and the variance of the distribution of person abilities; while the item parameter was the mean item difficulty (these person and item parameters are like those used in the previous research question). Since the original VERA dataset used in this study contained student response to only 48 test items, these items were used to simulate other datasets with 84 and 126 items (using the procedure already described in section 5.4.2 of this dissertation). Figure 5.7 and Figure 5.8 respectively describe the RMSE and the bias for the recovered person and item parameters across various test lengths, sample sizes and matrix booklet designs (The detailed values of the RMSE and bias from which Figure 5.7 and Figure 5.8 were respectively generated are presented in Tables 5.5 and Table 5.6).

The results show that test length and sample size are consistently related to the precision with which booklet designs recover person and item parameters of interest. Generally, it was found that increasing the sample size from 3000 did not lead to any significant gain in parameter recovery precision. Also, there was no bias in the recovery of parameters of interest ($0.00 \leq \text{Bias} \leq 0.02$) across all examined conditions. The mean person ability is very accurately recovered in all simulation conditions ($0.00 \leq \text{RMSE} \leq 0.03$). This implies that even with just 300 students, and with every student administered only six (out of total of 42 test items), the mean person ability can still be recovered very accurately and reliably. However, to accurately recover the variance of person ability (i.e., $\text{RMSE} \leq 0.04$), it is recommended that sample size be at least 3000 examinees, when using a design like Multi-matrix Booklet Design3. Conversely, to accurately recover the mean item location parameter, it is recommended that a minimum of 84 test items with a sample size of more than 3000 examinees be used, when applying multi-matrix booklet designs like Design1 or Design2.

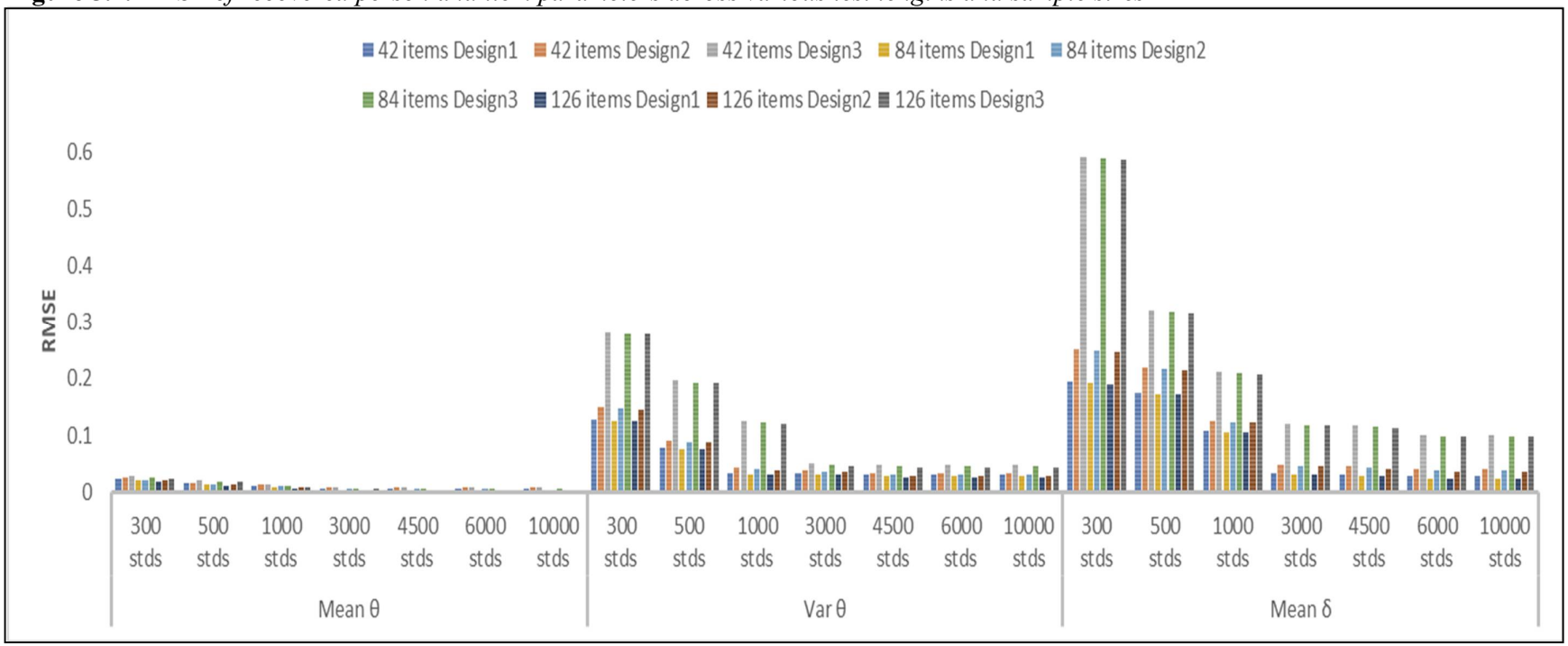
Similarly, the ANOVA results for the recovery of person and item parameters confirmed that sample size accounted for the major proportion of the variation in the log(RMSE) for the

recovered person and item parameters. Specifically, for recovery of the mean for the distribution of person abilities, sample size accounted for 87% of the total variation in the log(RMSE); while, for the variance of the distribution of person abilities and the mean item difficulty, it accounted respectively for 72% and 67% of the total variation, in the log(RMSE) of these parameters.

Also, matrix sparseness played an important role in explaining the variation in the log(RMSE) for recovering the mean item difficulty, accounting for 28% in the total variation. However, this effect was reduced when considering recovery of the mean and variance of the distribution of person ability, as matrix sparseness now accounted respectively for 6% and 19% of the total variation in the log(RMSE). Conversely, test length had a very small effect, accounting for less than 5% in the variation of the log(RMSE) for all person and item parameters.

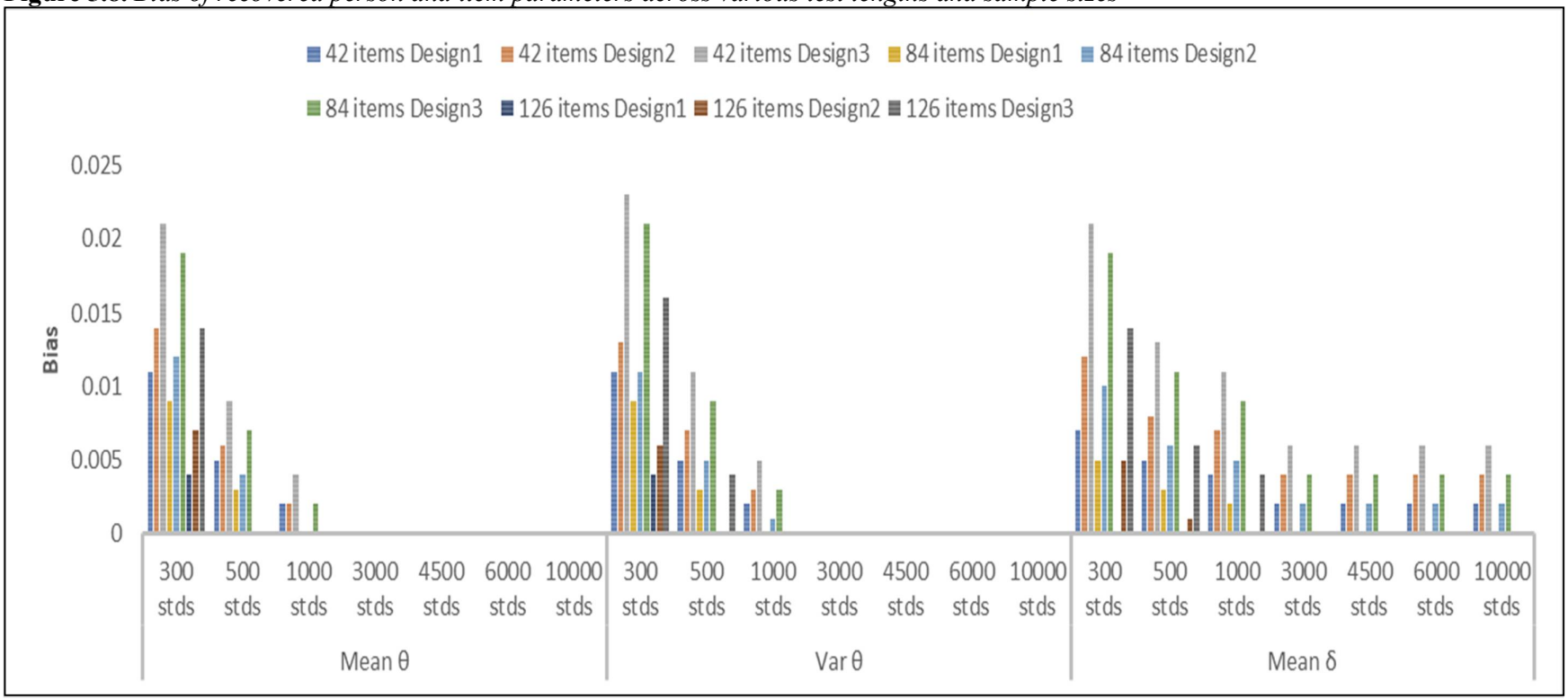
On the other hand, for recovery of the variance of the distribution of person abilities, the interaction between sample size and matrix sparseness had a small effect, accounting for 4% of the total variation in the log(RMSE). However, this interaction effect became even smaller when considering recovery of the mean for the distribution of person abilities and the mean item difficulty—where this interaction accounted for only 2% of the total variance, in the log(RMSE) for these recovered parameters. The details for all the above ANOVA results are presented in Table 5.7.

Figure 5.7. *RMSE of recovered person and item parameters across various test lengths and sample sizes*



Note. The abbreviation "stds" means "students". The first group of results are for the recovery of the mean person ability (i.e., mean θ); the second group of results, found at the centre, are for the recovery of the varaince of person abilities (i.e., var θ); while the last group of results are for the recovery of the mean item difficulty (i.e., mean δ). Each bar in the chart shows results for one simulation condition across 1000 replications (e.g., the first bar shows the RMSE for booklet Design1 when the test length is 42). Booklet Design1, Design2, and Design3 contain 57%, 71%, and 86% missing data respectively.

Figure 5.8. *Bias of recovered person and item parameters across various test lengths and sample sizes*



Note. The abbreviation "stds" means "students". The first group of results are for the recovery of the mean person ability (i.e., mean θ); the second group of results, found at the centre, are for the recovery of the varaince of person abilities (i.e., var θ); while the last group of results are for the recovery of the mean item difficulty (i.e., mean δ). Each bar in the chart shows results for one simulation condition across 1000 replications (e.g., the first bar shows the RMSE for booklet Design1 when the test length is 42). Booklet Design1, Design2, and Design3 contain 57%, 71%, and 86% missing data respectively.

Table 5.5. *RMSE of recovered person and item parameters across various test lengths and sample size*

		42 items			84 items			126 items		
	Sample size	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
Mean θ	300 stds	0.0224	0.0243	0.0281	0.0196	0.0215	0.0253	0.0185	0.0204	0.0242
	500 stds	0.0151	0.0157	0.0209	0.0123	0.0129	0.0181	0.0112	0.0118	0.017
	1000 stds	0.0101	0.0125	0.0125	0.0073	0.0097	0.0097	0.0062	0.0086	0.0086
	3000 stds	0.0064	0.0076	0.0088	0.0036	0.0048	0.006	0.0025	0.0037	0.0049
	4500 stds	0.0061	0.0072	0.0081	0.0033	0.0044	0.0053	0.0022	0.0033	0.0042
	6000 stds	0.0061	0.0072	0.0081	0.0033	0.0044	0.0053	0.0022	0.0033	0.0042
	10000 stds	0.0061	0.0072	0.0081	0.0033	0.0043	0.0053	0.0022	0.0032	0.0042
Var θ	300 stds	0.128	0.149	0.282	0.1252	0.1462	0.2792	0.1241	0.1451	0.2781
	500 stds	0.078	0.091	0.196	0.0752	0.0882	0.1932	0.0741	0.0871	0.1921
	1000 stds	0.034	0.043	0.124	0.0312	0.0402	0.1212	0.0301	0.0391	0.1201
	3000 stds	0.034	0.039	0.05	0.0312	0.0362	0.0472	0.0301	0.0351	0.0461
	4500 stds	0.03	0.033	0.048	0.0272	0.0302	0.0452	0.0261	0.0291	0.0441
	6000 stds	0.03	0.033	0.048	0.0272	0.0302	0.0452	0.0261	0.0291	0.0441
	10000 stds	0.03	0.033	0.047	0.0272	0.0302	0.0452	0.0261	0.0291	0.0431
Mean δ	300 stds	0.194	0.251	0.592	0.1912	0.2482	0.5892	0.1901	0.2471	0.5881
	500 stds	0.175	0.219	0.321	0.1722	0.2162	0.3182	0.1711	0.2151	0.3171
	1000 stds	0.108	0.126	0.211	0.1052	0.1232	0.2082	0.1041	0.1221	0.2071
	3000 stds	0.034	0.049	0.121	0.0312	0.0462	0.1181	0.0301	0.0451	0.1171
	4500 stds	0.031	0.045	0.117	0.0282	0.0422	0.1142	0.0271	0.0411	0.1131
	6000 stds	0.027	0.04	0.101	0.0242	0.0372	0.0982	0.0231	0.0361	0.0971
	10000 stds	0.027	0.04	0.101	0.0242	0.0372	0.0982	0.0231	0.0361	0.0971

Table 5.6. *Bias of recovered person and item parameters across various test lengths and sample sizes*

		42 items			84 items			126 items		
	Sample size	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
Mean θ	300 stds	0.011	0.014	0.021	0.009	0.012	0.019	0.004	0.007	0.014
	500 stds	0.005	0.006	0.009	0.003	0.004	0.007	0	0	0
	1000 stds	0.002	0.002	0.004	0	0	0.002	0	0	0
	3000 stds	0	0	0	0	0	0	0	0	0
	4500 stds	0	0	0	0	0	0	0	0	0
	6000 stds	0	0	0	0	0	0	0	0	0
	10000 stds	0	0	0	0	0	0	0	0	0
Var θ	300 stds	0.011	0.013	0.023	0.009	0.011	0.021	0.004	0.006	0.016
	500 stds	0.005	0.007	0.011	0.003	0.005	0.009	0	0	0.004
	1000 stds	0.002	0.003	0.005	0	0.001	0.003	0	0	0
	3000 stds	0	0	0	0	0	0	0	0	0
	4500 stds	0	0	0	0	0	0	0	0	0
	6000 stds	0	0	0	0	0	0	0	0	0
	10000 stds	0	0	0	0	0	0	0	0	0
Mean δ	300 stds	0.007	0.012	0.021	0.005	0.01	0.019	0	0.005	0.014
	500 stds	0.005	0.008	0.013	0.003	0.006	0.011	0	0.001	0.006
	1000 stds	0.004	0.007	0.011	0.002	0.005	0.009	0	0	0.004
	3000 stds	0.002	0.004	0.006	0	0.002	0.004	0	0	0
	4500 stds	0.002	0.004	0.006	0	0.002	0.004	0	0	0
	6000 stds	0.002	0.004	0.006	0	0.002	0.004	0	0	0
	10000 stds	0.002	0.004	0.006	0	0.002	0.004	0	0	0

Table 5.7. *Summary of ANOVA for recovery of person and item parameters considering sample size, test length, matrix sparseness (and the interaction between these factors).*

Source	d_f	log(RMSE) mean θ			log(RMSE) variance θ			log(RMSE) mean δ		
		Mean Square	F value	η^2	Mean Square	F value	η^2	Mean Square	F value	η^2
N	6	0.741	2585.67	.87	.741	86279.09	.72	1.136	36651.22	.67
L	2	0.684	1192.09	.02	.008	883.70	.04	.004	124.47	.02
S	2	0.319	556.27	.06	.600	65064.64	.19	1.334	43022.04	.28
N x L	12	0.143	41.58	.03	.001	26.63	.01	0.000	7.24	.01
N x S	12	0.035	10.13	.01	.019	2235.42	.04	0.018	572.82	.02
L x S	4	0.020	17.18	.00	.000	26.51	.00	0.000	12.43	.00
Residuals	24	0.007			.000			0.000		

Note. N = sample size (i.e., number of participating students); L = test length; and S = matrix sparseness (i.e., amount of missing data in the booklet design). N x L, N x S, and L x S represent interactions between these variables. The criterion variables were the log(RMSE) for the recovered mean of the distribution of person abilities; the log(RMSE) for the variance of the distribution of person abilities; and, the log(RMSE) for the mean of the recovered mean item difficulty. Further, $p < .001$ in all cases.

5.5.3 How efficiently are performance differences between policy relevant population subgroups recovered when using the various multi-matrix booklet designs (across conditions investigated in the study)?

Data were simulated for a population containing two population subgroups (Group1 and Group2). These groups were assumed to represent two policy relevant population subgroups (for instance, Group1 could represent high SES students, while Group2 represents low SES students).

Recovery of group difference in mean person ability

The results show that for a sample size of at least 1000 examinees, performance differences between population subgroups are recovered accurately and without bias ($0.015 \leq RMSE \leq 0.022$ and $0.000 \leq Bias \leq 0.002$). For sample sizes of less than 1000 examinees, unreliable estimates of performance differences between population subgroups are reported, especially when using a multi-matrix booklet design like Design3. Further, increasing the number of test participants improved the precision with which performance group differences were recovered. However, beyond a sample size of 3000 examinees, changes in sample size had a negligible effect on the recovery accuracy of the group difference in mean person ability. For instance, when using the sparsest multi-matrix booklet design and the case with the greatest difference in mean person ability between population subgroups, increasing the number of participating students from 3000 to 10,000 only resulted in an RMSE reduction of 0.0007 logits (i.e., an RMSE reduction from 0.0182 to 0.0175).

Further, the magnitude of the difference between the mean person ability between the population subgroups had a very negligible effect on the accuracy with which this parameter was recovered. For instance, considering the multi-matrix booklet Design1, the average reduction in the RMSE of the recovered group difference in mean person ability between population subgroups was 0.0006 logits when comparing the group with no performance difference (d=0) and the group with the greatest performance difference (d=2). Particularly, the major increase in the RMSE of the recovered group difference (in mean person ability) only occurred when using the sparsest multi-matrix design and when the sample size was less

than 1000 examinees. Thus, for a sample size of 300 students, RMSE of recovered group difference (in mean person ability) increased 0.012 logits, when comparing case with least performance difference between population subgroups (i.e., d=0) and case with the greatest performance difference between the population subgroups (i.e., d=2). A graphical representation of the results is displayed in Figure 5.9 below, while detailed values of the RMSE and bias across all investigated conditions are presented in Table 5.8.

Recovery of group difference in variance of person abilities

Generally, results for recovery of differences in variance of person ability between population subgroups were like those for the recovery of differences in mean person abilities between population subgroups ($0.019 \leq \text{RMSE} \leq 0.228$; $0 \leq \text{Bias} \leq 0.043$). However, the values of the RMSE and bias were larger than those when considering recovery of the group difference in mean person ability. The implication is that, in general, the group difference in mean person ability was more accurately recovered than the group difference in the variance of person ability.

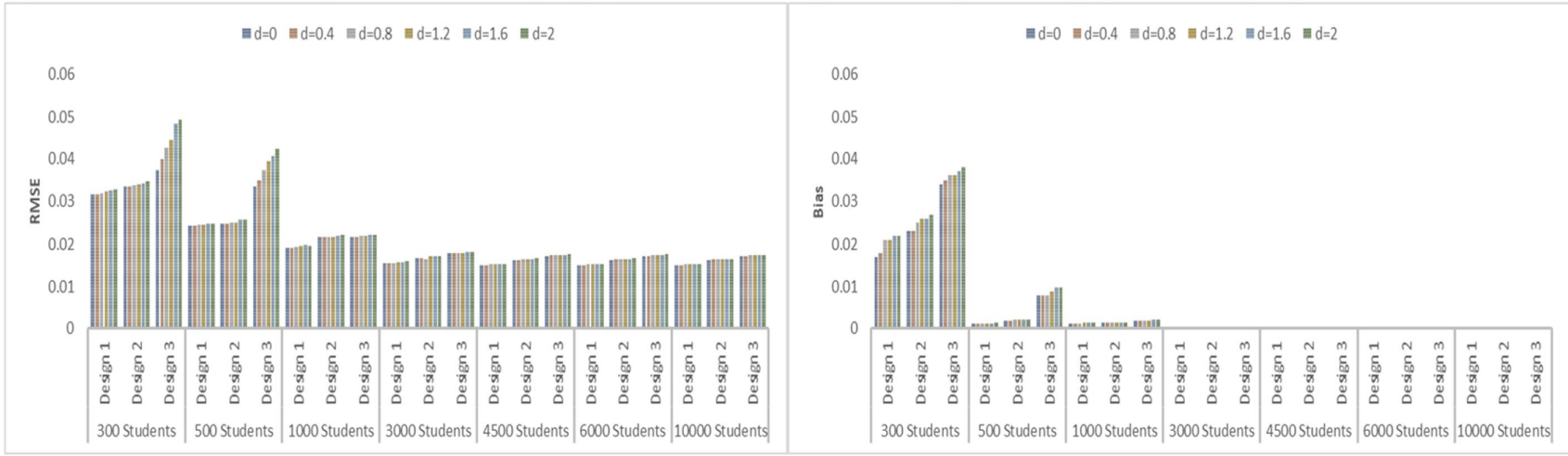
Further, for accurate recovery of the group difference in variance of person abilities, a minimum of 3000 test takers is required. This is unlike the case with recovering the group difference in mean person ability where the minimum requirement was 1000 test takers. Also, when the sample size became greater than 3000 test takers, further increments in sample size had negligible effects on the accuracy with which group difference in variance of person abilities were recovered. For instance, considering the sparsest multi-matrix design and the case with the greatest group difference in the variance of person abilities (i.e., d=2), increasing the sample size from 4500 test takers to 10,000 test takers only reduced the RMSE (of the recovered group difference in the variance of person ability) by 0.002 logits—that is, from 0.028 logits to 0.026 logits.

On the other hand, the magnitude of the group difference in the variance of person abilities had a negligible effect on how accurately this group difference was recovered. For instance, the average increase in the RMSE (of the recovered group difference in variance of person

ability) was 0.004, between the case where there was no group difference in the variance of person abilities (d=0) and the case where there was the greatest group difference in the variance of person abilities (d=2). Also, as sample size increased, the magnitude of the group difference (in the variance of person ability) had a smaller effect on the how accurately this person parameter was recovered.

Details of these results are presented in Figure 5.10 and on Table 5.9 below.

Figure 5.9. *RMSE and bias of recovered mean group difference in mean person ability between population subgroups*



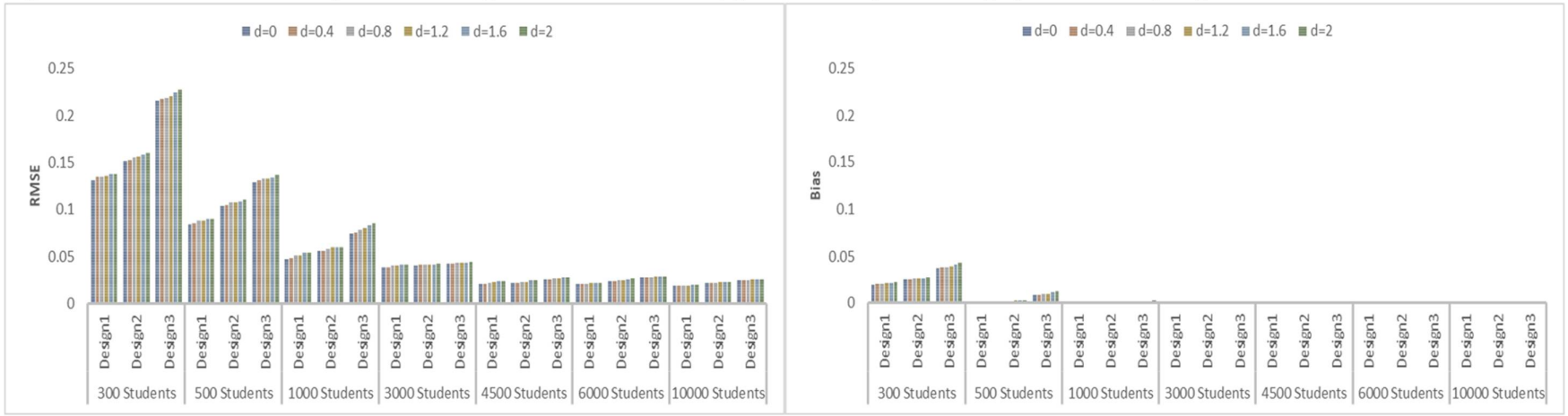
Note. BD1, BD2 and BD3 represent multi-matrix booklet Design1, Design2 and Design3 respectively. Further, d=0, d=0.4, …, d=2 represent the various degrees of difference in mean person ability between the population subgroups. For the condition d=0, there is no difference in mean person ability between population subgroups; while for d=2, the mean person ability for Group1 is -1 and the mean person ability for Group2 is 1. (N = 10,000 students, with each group containing 5,000 students).

Table 5.8. RMSE and bias of recovered group difference in mean person ability for population subgroups

	Condition	300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students			10000 Students		
		Design 1	Design 2	Design 3	Design 1	Design 2	Design 3	Design 1	Design 2	Design 3	Design 1	Design 2	Design 3	Design 1	Design 2	Design 3	Design 1	Design 2	Design 3	Design 1	Design 2	Design 3
RMSE	d=0	0.0315	0.0334	0.0372	0.0242	0.0248	0.0335	0.0192	0.0216	0.0216	0.0155	0.0167	0.0179	0.0152	0.0163	0.0172	0.0152	0.0163	0.0172	0.0152	0.0163	0.0172
	d=0.4	0.0316	0.0336	0.0398	0.0242	0.0248	0.0348	0.0192	0.0217	0.0218	0.0155	0.0168	0.0179	0.0152	0.0163	0.0174	0.0152	0.0164	0.0172	0.0152	0.0164	0.0172
	d=0.8	0.0319	0.0338	0.0426	0.0245	0.025	0.0372	0.0194	0.0217	0.0219	0.0156	0.0165	0.018	0.0153	0.0165	0.0174	0.0154	0.0164	0.0174	0.0154	0.0164	0.0174
	d=1.2	0.0322	0.034	0.0445	0.0246	0.025	0.0394	0.0195	0.0218	0.0219	0.0159	0.0171	0.018	0.0153	0.0166	0.0175	0.0153	0.0166	0.0174	0.0154	0.0164	0.0175
	d=1.6	0.0325	0.0343	0.0484	0.0247	0.0256	0.0406	0.0197	0.022	0.0221	0.0159	0.0171	0.0182	0.0153	0.0166	0.0175	0.0154	0.0166	0.0175	0.0154	0.0164	0.0175
	d=2	0.0328	0.0346	0.0492	0.0247	0.0257	0.0425	0.0196	0.0221	0.0222	0.016	0.0173	0.0182	0.0154	0.0168	0.0177	0.0154	0.0167	0.0177	0.0154	0.0165	0.0175
Bias	d=0	0.017	0.023	0.034	0.0011	0.0019	0.008	0.0012	0.0014	0.0018	0	0	0	0	0	0	0	0	0	0	0	0
	d=0.4	0.018	0.023	0.035	0.0011	0.0019	0.008	0.0012	0.0014	0.0018	0	0	0.0001	0	0	0	0	0	0	0	0	0
	d=0.8	0.021	0.025	0.036	0.0012	0.002	0.008	0.0012	0.0014	0.0019	0	0	0.0001	0	0	0	0	0	0	0	0	0
	d=1.2	0.021	0.026	0.036	0.0012	0.002	0.009	0.0013	0.0015	0.0019	0	0	0.0001	0	0	0	0	0	0	0	0	0
	d=1.6	0.022	0.026	0.037	0.0012	0.0021	0.01	0.0013	0.0015	0.002	0	0	0.0003	0	0	0.0001	0	0	0	0	0	0
	d=2	0.022	0.027	0.038	0.0013	0.0022	0.01	0.0013	0.0015	0.0021	0	0	0.0003	0	0	0.0001	0	0	0	0	0	0

Note. The conditions, d=0, d=0.4, …, d=2 represent the simulated difference in the mean person ability between the two population subgroups. d=2 implies the difference in the mean person ability between Group1 and Group2 was 2 logits (i.e., the mean person ability for students in Group1 was -1, while the mean person ability for students in Group2 was 1). Each result is pooled from 1000 replications.

Figure 5.10. *RMSE and bias of recovered difference in variance of person ability between population subgroups*



Note. d=0 to d=2 represent the group differnce in mean person ability between the population subgroups (N = 10,000 students, with each group containing 5,000 students).

Table 5.9. *RMSE and bias of the recovered group difference in variance of person ability for population subgroups*

		300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students			10000 Students		
	Condition	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
RMSE	d=0	0.131	0.151	0.216	0.084	0.104	0.129	0.048	0.056	0.075	0.039	0.041	0.043	0.021	0.022	0.026	0.021	0.024	0.028	0.019	0.022	0.025
	d=0.4	0.135	0.152	0.218	0.085	0.105	0.131	0.049	0.056	0.076	0.039	0.042	0.043	0.021	0.022	0.026	0.021	0.024	0.028	0.019	0.022	0.025
	d=0.8	0.135	0.155	0.219	0.088	0.108	0.133	0.052	0.058	0.079	0.041	0.042	0.044	0.022	0.023	0.027	0.021	0.025	0.028	0.019	0.022	0.025
	d=1.2	0.136	0.156	0.221	0.088	0.108	0.133	0.052	0.06	0.081	0.041	0.042	0.044	0.023	0.023	0.027	0.022	0.025	0.029	0.019	0.023	0.026
	d=1.6	0.138	0.158	0.225	0.09	0.109	0.134	0.054	0.06	0.083	0.042	0.042	0.044	0.024	0.025	0.028	0.022	0.026	0.029	0.02	0.023	0.026
	d=2	0.138	0.16	0.228	0.09	0.111	0.137	0.054	0.06	0.085	0.042	0.043	0.045	0.024	0.025	0.028	0.022	0.027	0.029	0.02	0.023	0.026
Bias	d=0	0.019	0.026	0.038	0.0013	0.0021	0.009	0.0014	0.0016	0.002	0	0.0001	0.0005	0	0	0	0	0	0	0	0	0
	d=0.4	0.02	0.026	0.039	0.0014	0.0022	0.009	0.0014	0.0016	0.002	0	0.0002	0.0005	0	0	0.0001	0	0	0	0	0	0
	d=0.8	0.02	0.027	0.039	0.0014	0.0022	0.01	0.0014	0.0017	0.0021	0	0.0002	0.0006	0	0	0.0001	0	0	0	0	0	0
	d=1.2	0.021	0.027	0.04	0.0015	0.0024	0.01	0.0015	0.0017	0.0022	0	0.0003	0.0006	0	0	0.0001	0	0	0	0	0	0
	d=1.6	0.021	0.027	0.041	0.0015	0.0024	0.011	0.0015	0.0017	0.0022	0.0001	0.0003	0.0007	0	0	0.0002	0	0	0	0	0	0
	d=2	0.022	0.028	0.043	0.0016	0.0024	0.012	0.0016	0.0018	0.0023	0.0001	0.0003	0.0007	0	0	0.0002	0	0	0	0	0	0

Note. The conditions, d=0, d=0.4, …, d=2 represent the simulated difference in the mean person ability between the two population subgroups. d=0.4 implies the difference in the mean person ability between Group1 and Group2 was 0.4 logits (i.e., the mean person ability for students in Group1 was -0.2, while the mean person ability for students in Group2 was 0.2). Each result is pooled from 1000 replications.

Chapter 6 Study II—Item-person match and parameter recovery efficiency

This chapter describes an extensive simulation study performed as the second major study of this dissertation project. The study examines how matching items to fit person abilities impacts on how well population and item parameters of interest are recovered. The chapter begins with a background on item and test information functions—a pertinent concept to understanding this research question, followed by a summary of some previous empirical studies on item-person match. Details on the study objectives, methodology and achieved results are then presented next.

6.1 Item and test information functions

The information function plays a key role in Item Response Theory, as it provides a means of precisely evaluating how well individual items in a test measure the level of a given latent trait—for instance, student ability, knowledge, or level of satisfaction (Zięba, 2013). Further, having *information* means knowing something about a specific topic or object. This is similar as in Statistics and Psychometrics (though more technical), where information is defined as the reciprocal of the variance with which a parameter could be estimated (Baker & Kim, 2017). Hence, estimating a parameter precisely (i.e., lesser variability) will imply more information will be known about the value of that parameter than if it was estimated with lesser precision (i.e. greater variability). Statistically, the degree of precision of an estimated

parameter is inversely proportional to the size of the variability of the estimates around the value of the parameter (Baker & Kim, 2017).

Importantly, IRT models offer a powerful technique for describing items and tests. This is also true for the selection of test items in cases where the IRT models are found to fit with the test data. This technique involves using *item information functions*. These (item information) functions play a crucial role in the development of tests, as they describe the contributions items make to the estimation of person abilities at given points along the ability continuum (Hambleton, Jones & Rogers, 1993). This contribution depends hugely on the item's discrimination power (with a greater value implying a steeper item characteristic curve and greater information provided by the item). Where exactly this contribution is made on the ability scale depends on the item's difficulty (Hambleton, 1989).

Further, summing the item information functions at every level that person ability is reported produces the test information function. Hence, the test information function is a measure of how much information is made available by all item responses on a test, concerning the latent trait or true score, θ (Johnson, 2018). Also, the precision with which ability is measured is greatly influenced by the amount of information provided by a test at a certain ability level—with more information provided resulting in more accurate estimates of person abilities. For the 1-PL IRT model (applied in this dissertation), the amount of information associated with any item, i, for a given ability level, θ, is given by the formula (Baker, 2001):

$$I_i(\theta) = P_i(\theta)Q_i(\theta), \qquad (6.1)$$

where $Q_i(\theta) = 1 - P_i(\theta)$.

Conversely, item information functions can be used to build tests which meet a desired set of statistical specifications. Birnbaum (1968) and Lord (1980) presented a procedure for doing this which involves using an item bank containing IRT item statistics and following these steps as outlined by Hambleton, Jones & Rogers (1993, p. 144):

1. Deciding what shape, the test information function will take (also known as the target information function).

2. Selecting item bank items that have item information functions which fill up hard-to-fill areas with respect to the target information function.
3. After adding every item to the test, computing the test information function for the chosen test items.
4. Proceeding with selection of test items until the test information function becomes approximately equal to the target information function.

This idea of matching test items to ability of test takers such that the items offer maximum information about test taker ability, forms the basis of techniques like test targeting and optimal item selection. In test targeting, this is done by (1) using background variables (which are related to student ability) to assign examinees test booklets having different mean difficulties, or (2) using ongoing examinee test performance to adaptively assign them to easy, fair or difficult subsequent test parts (Berger et al., 2019). Conversely, in optimal item selection, test developers choose items from an item bank based on how well they offer maximum information at a given point or range on the ability continuum (Hambleton & Jones, 1994). This is done by computer software using optimizing algorithms where test characteristics, like target information function and test length, are specified (Hambleton & Jones, 1994; van der Linden & Beokkooi-Timminga, 1989).

Thus, while the item information function describes how much statistical information a test item provides in estimating person abilities across the entire range of ability scores, the test information function describes how well the entire test performs in estimating ability across this ability continuum (Baker, 2001; Baker & Kim, 2017). Importantly, by varying the match between the distribution of person abilities and the distribution of item difficulties, the amount of information available for the estimation of IRT model parameters could be affected. As a result, this could also affect the accuracy with which item and population parameters (of interest) are recovered.

6.2 Empirical studies on item person-match

As noted by Svetina et al. (2013, p.336), "considerably less research has been conducted investigating IRT methods where a mismatch between item and person parameter distributions exist." Svetina and colleagues carried out a simulation study using the 1PL IRT model with short tests and small samples. They investigated estimation accuracy of item and person location parameters when the underlying item and person ability distributions were mismatched. Their results showed that the degree of mismatch between the item and person ability distributions influenced parameter recovery. However, the degree of mismatch likely to occur in practice has a relatively small effect on parameter recovery.

Further, Berger et al. (2019) performed another simulation study in which they compared the efficiency with which student ability was estimated using three item-person targeted designs. The designs used were the traditional targeted test design, the multistage test (MST) design, and the targeted multistage test (TMST) designs. They further investigated the degree to which the efficiency of these targeted designs was contingent on the correlation between (a) the ability-related background variables and the true examinee ability, and (b) examinee ability level and their classification into an ability group. Their results showed that examinee ability was generally more efficiently estimated with the targeted multistage design, especially when the ability-related background variable had a high correlation with true student ability. Also, targeted multistage testing resulted in efficient estimation of abilities for high- and low-ability students within the population.

As noted above, a dearth of research exists investigating IRT methods where a mismatch between item and person ability distributions exists. This is worse when looking at multi-matrix booklet designs. To my knowledge, no study has investigated the question of item-person mismatch and parameter recovery when using multi-matrix booklet designs[12]. Further, Svetina et al. (2013) used complete data in their study. Thus, it could be interesting to investigate what results are obtained when sparse datasets are used. Effects that are negligible

[12] Based on a google scholar search with the keywords: multi-matrix designs, person-item match, parameter recovery, item and person ability distribution match, IRT. This search was conducted in March 2019.

or small when using complete data, might become significant when using sparse data—as is typically the case with multi-matrix booklet designs. This might stem from the fact that less information is available for the estimation of IRT model parameters due to data sparseness.

6.3 Research objectives and research questions

In IRT measurement, person ability is measured most efficiently when items administered to examinees match their ability level on the measured latent trait (Lord, 1980; Rost, 2004; Berger et al. 2019). This is because more information is available for measurement when person ability matches item difficulty. This thus serves as the basis for targeted testing designs like multistage testing and computer adaptive testing (Berger et al., 2019). Also, several studies show that factors such as test length and sample size influence parameter recovery accuracy when carrying out measurement with IRT models (e.g., see DeMars, 2003; Swaminathan, Hambleton, Sireci, Xing and Rizavi, 2003; Yousfi, 2005; Wang and Chen, 2005; Chuah, Drasgrow and Leucht, 2006; He & Wheadon 2012). Further, a dearth of literature exists on item-person match and parameter recovery in IRT—worst still, when considering multi-matrix booklet designs.

Therefore, this study aims to investigate how the match between item and person ability distributions influence the efficiency of recovered person and item parameters when using sparse multiple matrix booklet designs. This will be achieved by specifically answering the following research questions:

1. Considering test length, how is the efficiency of recovered person and item parameters influenced by the match between item and person ability distributions in various sparse matrix booklet designs?

2. Considering sample size, how is the efficiency of recovered person and item parameters influenced by the match between item and person ability distributions in various sparse matrix booklet designs?

6.4 Data and procedure

To answer the above research questions, this study used real data, as well as simulated data. The real data was used to generate the simulated data. This was done such that simulated data has properties of the real dataset. The real dataset used was PISA 2012 Mathematics dataset for Germany. PISA (Programme for International Student Achievement) is a triennial international survey organized by the OECD (Organization for Economic Cooperation and Development) with the aim of evaluating education systems worldwide. It does this by testing the mathematics, science and reading competencies of 15-year-old students, who are towards the end of their compulsory education (OECD, 2018). The PISA 2012 dataset used in the study contains the cognitive responses of 5001 German students to 84 mathematics items (with a multi-matrix booklet design used in administering the items to the students). On the logit scale, the mean and standard deviation of student abilities were -0.02 and 1.281 respectively, while the mean and standard deviation of the item difficulties were 0.049 and 1.517 respectively. Respectively, these values on the PISA 2012 metric are 507.85 and 98.57 for the mean and standard deviation of person abilities, and 512.22 and 118.15 for the mean and standard deviation of item difficulties[13].

The simulated student response data were generated using the R package *irtoys* (Partchev, 2017). This was done by using (a) item difficulties from the real dataset and (b) the mean and standard deviation of person abilities from the real dataset. Five match conditions between item and person ability distributions were simulated as shown in Table 6.1 below.

For each of the above conditions, response data was simulated for test lengths of 42, 84, and 126 items. These numbers were chosen because they are plausible item numbers that could be used in operational test situations; and since the numbers are all multiples of 7, they can be

[13] The PISA scaled scores were got by applying the transformation (OECD, 2014): $\left[100\frac{L+0.0981}{1.2838}\right]+500$, where L represents the raw logit scores for the item or person parameter.

Table 6.1. *Match conditions between distributions of person ability and item difficulty in the simulated data*

Match condition	Mean Person Ability	Mean Item Difficulty
1 (d=0)	0	0
2 (d=0.2)	0.2	0
3 (d=0.4)	0.4	0
4 (d=0.6)	0.6	0
5 (d=0.8)	0.8	0

Note. In condition 1 (d=0), there is perfect match between the person ability distribution and the distribution of item difficulties (Both distributions having a mean of 0). The mean person ability and the mean item difficulty are given in logits.

used in creating BIB7 booklet designs[14]. Importantly, BIB7 booklet designs can only be created when the test length is a multiple of 7 (See Table 6.2 for a structural overview of the experimental conditions and variables used in this study).

Further, to learn how sample size and person-item match influence parameter recovery efficiency, sample sizes of 300, 500, 1,000, 3,000, 4,500, and 6,000 students were simulated for each of the match conditions as described in Table 6.1 above. These numbers were chosen because they cover a wide spectrum of the possible number of students that can partake in an operational test (e.g., PISA requires a minimum sample size of 4,500 students from each participating country). All experimental conditions were fully crossed. Doing so, yielded 4 (booklet designs[15]) x 6 (sample sizes) x 3 (total number of items) x 5 (Person-Item match conditions) = 360 experimental conditions. Further, to ensure the stability of results, each experimental condition was replicated 1000 times. Thus, in this study a total number of 360 * 1000 = 360,000 datasets were analysed.

[14] As was already explained in the introductory chapter of the PhD dissertation, BIBD7 designs have many desirable characteristics (e.g., every item pair occurs an equal number of times, and does so at least once). Also, the BIBD design and several variants of it are used in many LSAs like PISA (Gonzalez and Rutkowski, 2010).
[15] The booklet designs used were the same as those used in Study I of this dissertation (See Section 5.3—Particularly, Figure 5.1 on page 58, gives an overview of these booklet designs).

Table 6.2. *Overview of the study design*

Data to be studied	Experimental Conditions				Major Dependent Variables
	Booklet Design	Sample Size	Total Number of Items	Person-Item match condition	
Simulated Data	1,2,3,4	300, 500, 1,000, 3,000, 4,500, 6,000,	42, 84, 126	1, 2, 3, 4, 5	Bias & RMSE for item and person parameters at the population level

Note. The booklet design (see Section 5.4.3) determines the number of items per student, i.e., the matrix sparseness. All experimental conditions are fully crossed. In this study, the simulated data were created using item and person characteristics from the PISA 2012 mathematics dataset for Germany. The person parameters investigated were the mean person ability and the variance of person abilities, while the item parameter examined was the mean item difficulty. The person match condition describes the match between the distribution of the person abilities and the distribution of the item difficulties (See Table 6.1)

6.5 Method of data analysis

The method of data analyses was the same as in Study I of this dissertation (See Section 5.4). Importantly, inferential statistics were reported following very closely, the recommendations of Harwell, Stone, Hsu, & Kirisci (1996) and Feinberg & Rubright (2016), for reporting results of simulation studies performed in Item Response Theory and Psychometrics. Thus, the simulation results were "summarised as ANOVA's to highlight the main effects" (See, Feinberg & Rubright, 2016, p. 44). Also, as recommended by Harwell et al. (1996, p.21), a non-linear, log transformation of the dependent variable (i.e., RMSE) was performed to increase the likelihood of the normality assumption being satisfied. Particularly, since the item difficulties were normally distributed, it was expected that the log(RMSE) of parameters computed from these item parameter ought also be asymptotically normally distributed, with a mean and variance depending on the number of replications (Bartlett & Kendall, 1946; Harwell et al., 1996). In the same manner, effect sizes of independent variables were computed using eta squared. As emphasized by Levine & Hullett (2002, p. 612), "Eta squared

(η^2) is the most commonly reported estimate of effect size for the ANOVA." Further, it is easy to interpret, as it represents the percentage of variance associated with each independent variable; and all sources of variation (with their individual errors) add up to 1.00.

6.6 Results and discussion

The results of how item-person match relates to efficiency with which person and item parameters are recovered are described below under three sub-sections. Each sub-section will describe how item-person match impacts the recovery of a single person or item parameter. At the end of the last section, a brief summary is given on how test length, sample size and item-person match relate to the efficiency with which item and person parameters are recovered.

6.6.1 Item-person match and efficiency of mean person ability estimate recovery

Table 6.4 and Table 6.5 respectively give detailed values of the RMSE and bias (for recovered means of the distribution of person abilities) across all conditions of item-person match, sample size and test length as was investigated in this study. The results show that, for a sample size of more than 1000 students, item-person match has almost no effect on the efficiency[16] with which the mean person ability is recovered. However, even when the sample size is less than 1000, the effect of item-person match on both the RMSE and bias is still extremely small. For instance, for a sample size of 500 students, test length of 42 items, and considering the sparsest multi-matrix design, the RMSE of the recovered mean person ability increased only 0.002 logits (i.e., from 0.021 to 0.023) when comparing the perfectly matched condition and the most mismatched condition. Similarly, for a sample size of 300 students, test length of 42 items, and the sparsest multi-matrix design, the RMSE of the recovered mean person ability increased 0.005 logits (i.e., from 0.028 to 0.033), when the

[16] Efficiency here has to do with how precisely the various sparse booklet designs recover the true parameter value.

perfectly matched condition and the most mismatched conditions were considered respectively (See Table 6.4 for details).

For a graphical overview of these results, Figure 6.2 displays values of the RMSE for the recovered mean of the distribution of person abilities across the various experimental conditions investigated. Importantly, although the mean of the person ability distribution was recovered accurately in all simulation conditions, the RMSEs became slightly greater as items and persons became more mismatched. Figure 6.3 shows the distribution of recovered mean person abilities across 1000 replications, when test length is 42 items. A simplified version of the same results is presented in Appendix A.3. Further, irrespective of the degree of match between the distributions of item difficulties and person abilities, no bias was found in the recovery of the mean person ability. The bias in all cases examined was always less than 0.03 (See Table 6.5 for details).

On the other hand, the ANOVA results showed that amongst all factors investigated, sample size accounted for almost all the variation in the log(RMSE) for the recovered mean in the distribution of person abilities, accounting for up to 93% of the total variance. Matrix sparseness and test length had small effects, accounting for 4% and 1% respectively of the total log(RMSE) variation. Further, distribution match contributed for less than 1% of this total variation, same as the various interaction effects (i.e., sample size x test length, and sample size x matrix sparseness). Similarly, there was no significant interaction effect between test length and matrix sparseness. The complete results from these ANOVA analyses are presented in Table 6.6.

Thus, in conclusion, irrespective of how sparse the booklet design was, the match between the distribution of person abilities and the distribution of item difficulties had a negligible effect on the precision with which the mean of the person ability distribution was recovered (across all experimental conditions investigated in the study).

Table 6.4. *RMSE of recovered mean person ability across various levels of item-person match, sample size and test length*

	Match	300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students		
		Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
42 Items	d=0	0.0224	0.0243	0.0281	0.0151	0.0157	0.0209	0.0101	0.0125	0.0125	0.0064	0.0076	0.0088	0.0061	0.0072	0.0081	0.0062	0.0072	0.0083
	d=0.2	0.0236	0.0259	0.0294	0.0153	0.0159	0.0211	0.0103	0.0129	0.0129	0.0065	0.0078	0.0087	0.0061	0.0073	0.0080	0.0062	0.0073	0.0084
	d=0.4	0.0243	0.0271	0.0311	0.0157	0.0164	0.0212	0.0109	0.0136	0.0136	0.0067	0.0080	0.0090	0.0063	0.0075	0.0083	0.0064	0.0075	0.0085
	d=0.6	0.0255	0.0283	0.0322	0.0161	0.0172	0.0219	0.0108	0.0138	0.0138	0.0069	0.0079	0.0094	0.0065	0.0074	0.0087	0.0066	0.0074	0.0089
	d=0.8	0.0274	0.0296	0.0330	0.0172	0.0184	0.0226	0.0116	0.0141	0.0141	0.0071	0.0081	0.0096	0.0067	0.0076	0.0089	0.0068	0.0076	0.0090
84 Items	d=0	0.0210	0.0243	0.0291	0.0143	0.0149	0.0191	0.0100	0.0122	0.0131	0.0063	0.0074	0.0082	0.0060	0.0068	0.0075	0.0060	0.0068	0.0075
	d=0.2	0.0224	0.0243	0.0293	0.0143	0.0152	0.0202	0.0102	0.0121	0.0132	0.0066	0.0073	0.0083	0.0062	0.0068	0.0076	0.0062	0.0068	0.0076
	d=0.4	0.0227	0.0245	0.0296	0.0145	0.0151	0.0202	0.0100	0.0123	0.0132	0.0067	0.0076	0.0083	0.0063	0.0071	0.0076	0.0062	0.0070	0.0077
	d=0.6	0.0231	0.0244	0.0297	0.0148	0.0152	0.0201	0.0105	0.0125	0.0135	0.0071	0.0080	0.0086	0.0067	0.0075	0.0079	0.0066	0.0074	0.0079
	d=0.8	0.0236	0.0247	0.0297	0.0152	0.0154	0.0203	0.0106	0.0128	0.0137	0.0071	0.0081	0.0088	0.0067	0.0076	0.0081	0.0067	0.0076	0.0081
126 Items	d=0	0.0194	0.0199	0.0221	0.0139	0.0142	0.0185	0.0100	0.0121	0.0130	0.0059	0.0066	0.0076	0.0056	0.0062	0.0071	0.0054	0.0062	0.0071
	d=0.4	0.0195	0.0199	0.0227	0.0139	0.0141	0.0186	0.0100	0.0123	0.0131	0.0058	0.0064	0.0077	0.0057	0.0069	0.0070	0.0054	0.0069	0.0070
	d=0.8	0.0195	0.0201	0.0229	0.0137	0.0143	0.0185	0.0103	0.0123	0.0134	0.0058	0.0064	0.0080	0.0057	0.0069	0.0073	0.0055	0.0069	0.0073
	d=1.2	0.0198	0.0203	0.0231	0.0140	0.0145	0.0187	0.0106	0.0125	0.0132	0.0055	0.0065	0.0082	0.0058	0.0071	0.0075	0.0057	0.0070	0.0074
	d=1.6	0.0198	0.0205	0.0233	0.0140	0.0145	0.0189	0.0106	0.0124	0.0133	0.0056	0.0068	0.0082	0.0060	0.0073	0.0075	0.0060	0.0073	0.0075

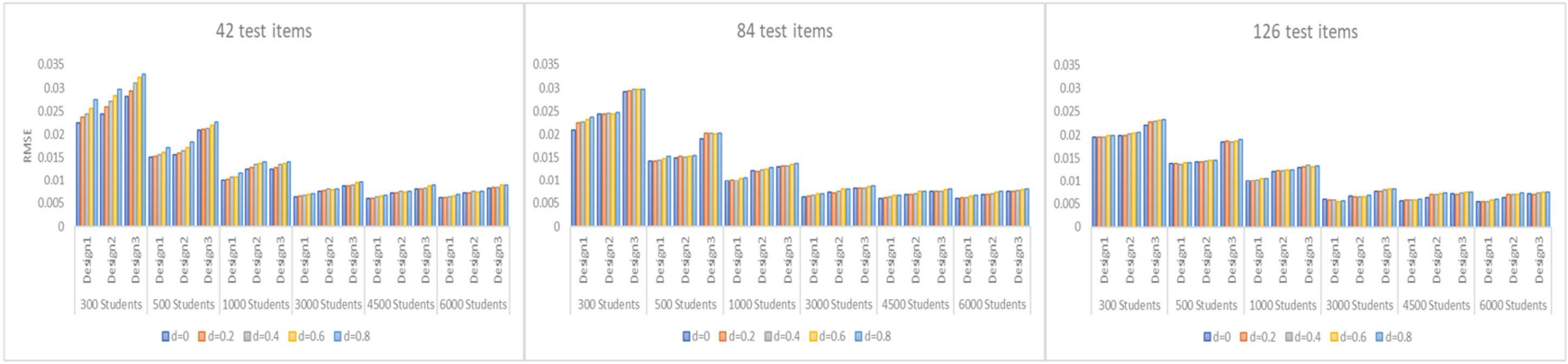
Note. "Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all conditions of distribution match investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities have means of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

Table 6.5. *Bias of recovered mean person ability across various levels of item-person match, sample size and test length*

	Match	300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students		
		Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
42 Items	d=0	0.011	0.014	0.021	0.005	0.006	0.009	0.002	0.002	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.012	0.015	0.021	0.005	0.006	0.009	0.002	0.002	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.014	0.016	0.023	0.005	0.006	0.009	0.002	0.003	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.015	0.016	0.023	0.006	0.006	0.010	0.002	0.002	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.8	0.015	0.017	0.025	0.006	0.007	0.010	0.003	0.003	0.005	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
84 Items	d=0	0.008	0.012	0.018	0.004	0.005	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.008	0.013	0.020	0.004	0.005	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.010	0.013	0.021	0.005	0.005	0.010	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.010	0.014	0.021	0.005	0.006	0.010	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.8	0.011	0.014	0.022	0.006	0.006	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
126 Items	d=0	0.006	0.009	0.015	0.002	0.003	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.006	0.009	0.016	0.002	0.003	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.006	0.010	0.017	0.002	0.004	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.007	0.010	0.017	0.003	0.004	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.8	0.007	0.011	0.018	0.003	0.005	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

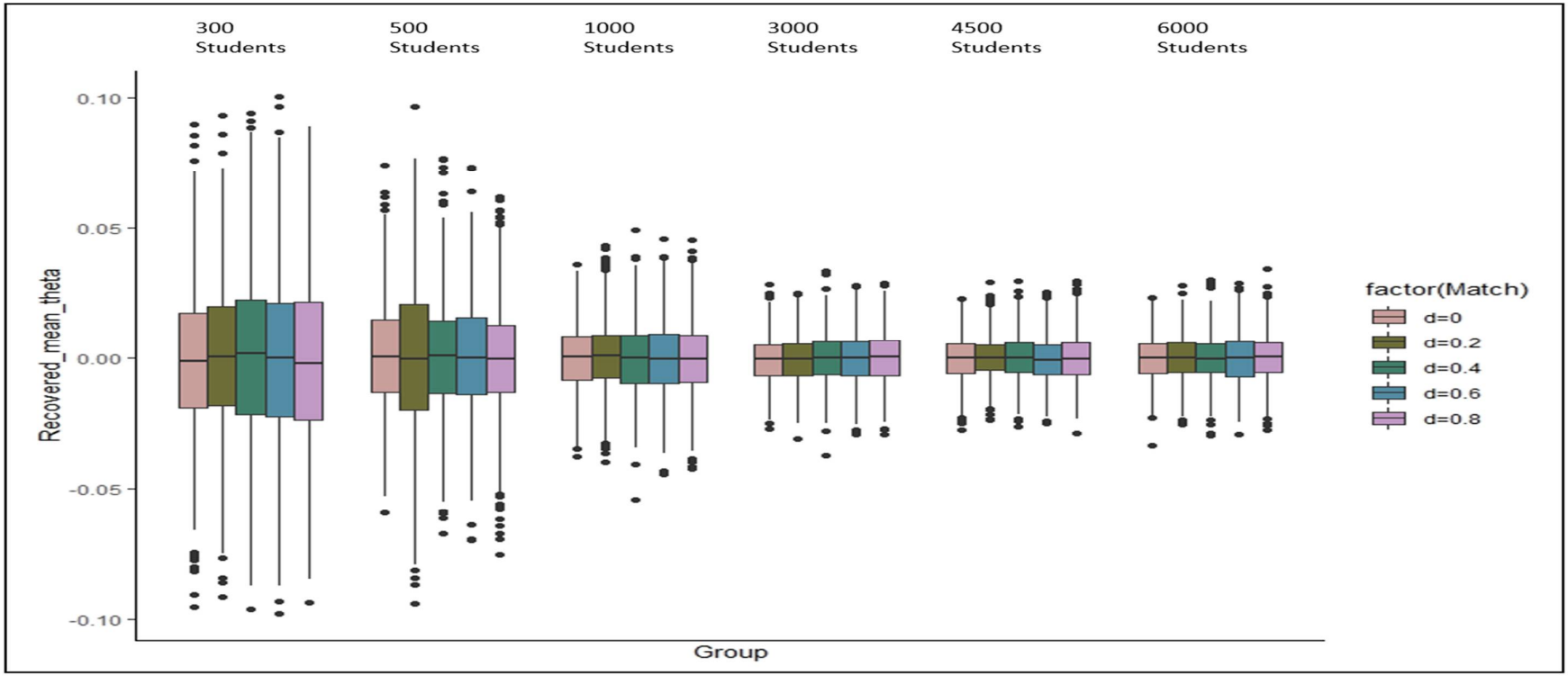
Note. "Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all conditions of distribution match investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities have means of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

Figure 6.2. *RMSE for the recovery of the mean person ability estimate across all experimental conditions*



Note. The bar plots are in groups of 3's for each sample size (i.e., number of students). For instance, in the first panel which is a bar plot representing the case for a test length of 42 items, the first three groups are results for a sample size of 300 students. Further, each of these three groups represents results for one multi-matrix design. The multi-matrix designs are labelled D1, D2 and D3, with the designs becoming sparser moving from D1 to D3 (Multi-matrix Design D1 contains 57% missing data, while designs D2 and D3 contain 71% and 86% missing data respectively). Also, "Distribution Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). For the conditions d=0.2, d=0.4, d=0.6 and d=0.8, the mean for the distribution of person abilities is 0.2, 0.4, 0.6 and 0.8 respectively. In all distribution match conditions, the mean item difficulty is fixed at 0.

Figure 6.3. *Distribution of recovered mean person ability estimate for different item-person match conditions (Case for sparsest multi-matrix design and test length of 42)*



Note. "Match" represents the degree of overlap between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect overlap between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all conditions of distribution match investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively.

Table 6.6: *ANOVA results with log(RMSE) of the mean for the distribution of person abilities being the dependent variable.*

Source	df	Mean square	F value	η^2
Sample Size	5	2.3474	15509.25	.93
Test Length	2	0.0856	565.78	.01
Matrix Sparseness	2	0.2758	1828.69	.04
Distribution Match	4	0.0104	68.44	.00
Sample Size x Test Length	10	0.0050	33.11	.00
Sample Size x Matrix Sparseness	10	0.0049	32.22	.00
Test Length x Matrix Sparseness	4	0.0004	2.38	.00
Residuals	232	0.0002		

Note. $p < .05$ for all conditions, except for the Test Length x Matrix Sparseness interaction.

6.6.2 Item-person match and efficiency of variance (of person abilities) recovery

Table 6.7 and Table 6.8 show details of the RMSE and the bias for the recovered variance of person abilities across various levels of item-person match, sample sizes and test lengths—as investigated in the study. In general, the results for the recovery of the variance of person abilities were like those for the recovery of the mean person ability. There was no bias in all conditions investigated, irrespective of how mismatched the person ability distributions and the item difficulty distributions were (Bias < 0.03 in all examined conditions, see Table 6.8 for details). Figure 6.4 describes how item-person match is related to the efficiency with which the variance of person abilities is recovered considering various multi-matrix designs, sample sizes and test lengths.

The major difference in the recovery of both parameters (i.e., the mean of the distribution of person abilities and the variance of the distribution of person abilities) was that values of the RMSEs for the recovered variance of distribution of person abilities were generally larger. While the RMSE and bias for the recovered mean of person ability distribution lay in the range [0.005, 0.033] and [0.000, 0.025] respectively; these values for recovery of variance of person abilities were in the range [0.013, 0.318] and [0.000, 0.027] for the RMSE and bias respectively.

More specifically, the increase in the RMSEs as a result of increasing item-person mismatch was very small. For instance, for a test length of 42 items, the average increase in RMSE from the perfectly matched condition (d=0) to the most mismatched condition (d=0.8) was 0.034. However, this value became even smaller as test length increased—thus, for the 84- and 124-itemed tests, this increase was 0.027 and 0.016 respectively. This implies that the impact of item-person mismatch, (though small), grew even smaller as test length increased. Doubling the test length from 42 items to 84 items resulted in a reduction of about 21% in the RMSE between the perfectly matched condition (d=0) and the most mismatched condition (d=0.8).

Similarly, there was an interaction between “Sample Size and Matrix Sparseness”. Thus, increasing the sample size reduced the effect of matrix sparseness on the precision with which the variance of the distribution of person abilities was recovered (i.e., smaller RMSE). Taking the sparsest multiple matrix design with a test length of 42 as an example, the increase in RMSE from the perfectly matched case (d=0) to the most mismatched case (d=0.8) was 0.036 when the sample size was 300 students; however, when the sample size was now 6000 students, the increase in RMSE from the perfectly matched case (d=0) to the most mismatched case (d=0.8) became 0.026. This means that there was a 10% reduction in the effect of item-person mismatch when the sample size was increased from 300 to 6000 students.

Table 6.9 summarizes ANOVA results for the recovered variance of the distribution of person abilities and factors investigated in the study (i.e., number of participating students, test length, matrix sparseness in booklet design, match between the distribution of person abilities and the distribution of item difficulties; and, the interactions between all of these factors). As was the case with recovering means of the distribution of person abilities, the major source of variance in the log(RMSE) for the recovered variance of the distribution of person abilities was the number of participating students (accounting for 65% of the total variance).

Further, the degree of sparseness in the booklet design accounted for 15% of the total variance in the log(RMSE); while test length and distribution match (i.e. the degree of overlap between the distribution of person abilities and the distribution of item difficulties) accounted for 8% and 15% respectively. Importantly, the interactions (Sample Size x Test Length; and, Sample Size x Matrix Sparseness) had a greater effect than was the case when considering recovery of the mean for the distribution of person abilities, with these interactions accounting for 2% and 4% of the total variation respectively. On the other hand, the interaction between test length and matrix sparseness accounted for less than 1% in the total variation of the log(RMSE).

Table 6.7. *RMSE of recovered variance of person abilities across various levels of item-person match, sample size and test length*

	Match	300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students		
		Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
42 Items	d=0	0.128	0.149	0.282	0.078	0.091	0.196	0.034	0.043	0.124	0.034	0.039	0.05	0.030	0.033	0.048	0.030	0.033	0.048
	d=0.2	0.139	0.157	0.291	0.089	0.100	0.205	0.042	0.051	0.135	0.043	0.047	0.058	0.037	0.041	0.056	0.036	0.040	0.055
	d=0.4	0.145	0.167	0.303	0.097	0.108	0.214	0.053	0.062	0.142	0.051	0.056	0.069	0.048	0.051	0.066	0.044	0.048	0.063
	d=0.6	0.152	0.175	0.312	0.105	0.116	0.223	0.061	0.070	0.150	0.061	0.066	0.077	0.056	0.060	0.074	0.049	0.054	0.069
	d=0.8	0.164	0.183	0.318	0.116	0.127	0.232	0.069	0.079	0.161	0.069	0.074	0.084	0.066	0.068	0.084	0.057	0.062	0.074
84 Items	d=0	0.115	0.129	0.257	0.076	0.081	0.161	0.023	0.031	0.120	0.025	0.031	0.047	0.018	0.025	0.038	0.018	0.025	0.028
	d=0.2	0.121	0.135	0.264	0.082	0.088	0.167	0.031	0.038	0.128	0.031	0.038	0.055	0.025	0.031	0.046	0.024	0.032	0.025
	d=0.4	0.129	0.143	0.270	0.090	0.094	0.175	0.037	0.046	0.134	0.039	0.044	0.061	0.033	0.039	0.052	0.032	0.038	0.031
	d=0.6	0.135	0.148	0.278	0.096	0.102	0.181	0.045	0.052	0.142	0.045	0.052	0.067	0.039	0.045	0.06	0.038	0.046	0.039
	d=0.8	0.143	0.155	0.284	0.104	0.108	0.190	0.051	0.058	0.148	0.053	0.058	0.075	0.045	0.053	0.066	0.046	0.052	0.045
126 Items	d=0	0.115	0.127	0.234	0.067	0.078	0.135	0.018	0.021	0.115	0.015	0.023	0.037	0.013	0.017	0.025	0.013	0.017	0.024
	d=0.2	0.121	0.132	0.239	0.071	0.081	0.141	0.022	0.026	0.121	0.020	0.028	0.038	0.014	0.022	0.029	0.014	0.021	0.028
	d=0.4	0.125	0.136	0.245	0.077	0.088	0.145	0.028	0.031	0.125	0.024	0.032	0.038	0.017	0.025	0.034	0.014	0.024	0.031
	d=0.6	0.131	0.142	0.249	0.081	0.091	0.151	0.033	0.035	0.131	0.030	0.038	0.040	0.019	0.029	0.037	0.015	0.025	0.032
	d=0.8	0.135	0.147	0.255	0.087	0.097	0.155	0.038	0.041	0.137	0.032	0.041	0.043	0.023	0.031	0.041	0.017	0.027	0.032

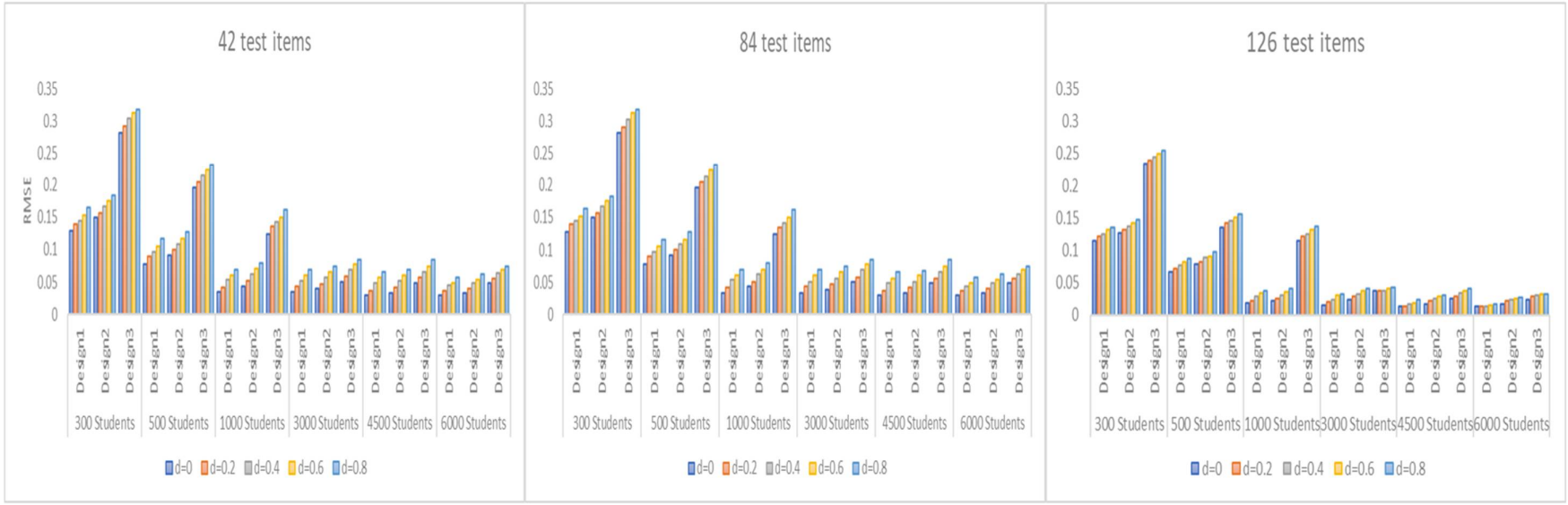
Note. "Distribution Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all conditions of distribution match investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

Table 6.8. *Bias of recovered variance of person abilities across various levels of item-person match, sample size and test length*

	Match	300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students		
		Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
42 Items	d=0	0.011	0.013	0.023	0.005	0.007	0.011	0.002	0.003	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.011	0.014	0.024	0.005	0.008	0.012	0.002	0.003	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.012	0.014	0.025	0.006	0.008	0.013	0.002	0.004	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.013	0.014	0.025	0.006	0.009	0.013	0.002	0.005	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.8	0.013	0.015	0.027	0.006	0.01	0.014	0.003	0.005	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
84 Items	d=0	0.008	0.011	0.018	0.003	0.005	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.008	0.012	0.018	0.003	0.006	0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.008	0.012	0.02	0.004	0.006	0.012	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.009	0.013	0.02	0.005	0.006	0.012	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.8	0.009	0.014	0.022	0.005	0.007	0.015	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
126 Items	d=0	0.005	0.006	0.009	0.000	0.001	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.005	0.007	0.010	0.001	0.001	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.006	0.008	0.012	0.001	0.002	0.006	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.006	0.008	0.013	0.001	0.003	0.007	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.8	0.007	0.009	0.015	0.002	0.003	0.009	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note. "Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all the match conditions investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

Figure 6.4. *RMSE for the recovery of the variance of person abilities for different item-person match conditions*



Note. The multi-matrix booklet designs become sparser moving from Design1 to Design3. The multi-matrix design Design1 represents the design with 57% missing data, while Desing2 and Design3 are multi-matrix de signs with 71% and 86% of missing data respectively. d=0, d=0.2, ..., d=0.8 represent the distribution match (i.e., the degree of match between the distribution of item difficulties and the distribution of person abilities). Moving from d=0 to d=0.8, the two distributions (of item difficulties and person abilities) become more misaligned with d=0 being the condition of perfect match between both distributions, and d=0.8 being the most mismatched condition

Table 6.9. *ANOVA results with log(RMSE) of recovered variance for the distribution of person abilities being the criterion.*

Source	df	Mean square	F value	η^2
Sample Size	5	3.186	1483.72	.65
Test Length	2	1.156	449.31	.08
Matrix Sparseness	2	2.154	837.34	.15
Distribution Match	4	0.340	132.06	.05
Sample Size x Test Length	10	0.052	20.19	.02
Sample Size x Matrix Sparseness	10	0.113	43.84	.04
Test Length x Matrix Sparseness	4	0.020	7.90	.00
Residuals	232	0.0002		

Note. $p < .001$ for all cases.

6.6.3 Item-person match and recovery of mean item difficulty

Table 6.10 and Table 6.11 respectively show the RMSE and Bias for the recovered mean item difficulty across the various conditions of item-person match, sample sizes and test lengths investigated in the study. These results also follow a similar pattern to the above results for the recovery of the mean person ability and variance of person ability (though the values of the RMSE for the recovered mean item difficulty were slightly larger, $0.013 \leq$ RMSE ≤ 0.429). Also, item-person match had no effect on the recovery of the mean item difficulty except when the sample size was less than 1000 students. That notwithstanding, this effect was also very small (See Table 6.10 for details). Further, though there was no bias in recovering mean item difficulty across all conditions of item person match, sample size, and test length ($0.000 \leq$ Bias ≤ 0.025); accurate recovery of this parameter especially with the sparsest booklet design required more than 1000 students.

On the other hand, upon further examination of the recovery of item difficulties at an individual item level, the results showed that the nature of item-person match affected recovery precision for individual items in completely different ways. For instance, shifting the distribution of person abilities to the right, by using students with higher abilities (the distribution of item difficulties kept fixed, with a mean of 0), resulted in the item difficulty for the difficult items being recovered more accurately than the item difficulties for the easier items. As an example, for the distribution match condition, d=0.8, where the distribution of person abilities had a mean of 0.8, while the mean of the distribution of item difficulties remained fixed at 0; the results showed that the more difficult items were recovered more accurately than the easier items. Conversely, for the distribution match condition, d=0, where both distributions of item difficulties and person abilities were perfectly aligned (with both distributions having a mean of 0), the results showed that the item difficulties for all items were recovered with almost the same degree of accuracy, except for a few extremely easy and extremely difficult items (See Figure 6.6 and Figure 6.7 for details).

To further verify these results, another distribution match condition, d= -0.8, was investigated, in which the distribution of person abilities was now shifted to the left, so that

the mean person ability was -0.8 (with the distribution of item difficulties still kept fixed, with a mean of 0). In this scenario, shifting the distribution of person abilities to the left resulted in the item difficulties, for the easier items, being recovered more accurately than item difficulties for the more difficult items (See Figure 6.6, Figure 6.7 and Figure 6.8 for details of these results).

Similarly, ANOVA analyses were performed to examine how the various investigated factors accounted for variation in the log(RMSE) for the recovered mean item difficulty. A summary of these results is presented in Table 6.12. Further, as was the case for the recovery of the mean and variance for the distribution of person abilities, sample size accounted for a major part of the total variation (accounting for 70% of the variation in the log(RMSE) of the recovered mean item difficulty). However, the degree of sparseness in the booklet design also played a crucial role, accounting for 22% of the total variation in the RMSE, which was more than five times the magnitude of the variance accounted for by the same factor when considering recovery of the mean of the distribution of person abilities. Test length and distribution match accounted for 2% and 1% respectively; while the interaction between number of participating students (sample size) and matrix sparseness in booklet design had a larger impact, accounting for 3% percent of the total variance.

To summarize the above results on how person-item match affected the recovery of person and item parameters investigated, it was found that: (1) When considering recovery of the mean person ability, the match between item and person ability distributions had a negligible effect on the precision with which item and person parameters were recovered—with the recovery precision becoming only slightly worse as the two distributions became more misaligned; (2) As test length increased, the effect of a mismatch (between the distribution of item difficulties and the distribution of person abilities) on person and item parameter recovery precision, became more negligible. For instance, for a 42-item test, there was an average reduction of 0.0018 in the RMSE of the recovered mean person ability between the perfectly matched case (with mean person ability of 0, mean item difficulty of 0), and the most mismatched case (where the mean person ability was 0.8, and the mean item difficulty was 0). However, when test length was increased to 84-items, this average reduction dropped

to 0.0008. Also, when the test length was further increased to 126 items, this average reduction in the RMSE of the recovered mean person ability became 0.0005. Thus, as test length increased, the effect of a mismatch between the distribution of person abilities and the distribution of item difficulties, grew more negligible.

Table 6.10. *RMSE of recovered mean item difficulty across various levels of item-person match, sample size and test length*

	Match	300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students		
		Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
42 Items	d=0	0.194	0.251	0.392	0.175	0.219	0.321	0.108	0.126	0.211	0.034	0.049	0.121	0.031	0.045	0.117	0.027	0.04	0.101
	d=0.2	0.200	0.257	0.402	0.18	0.224	0.327	0.113	0.132	0.217	0.039	0.053	0.124	0.036	0.05	0.122	0.03	0.044	0.105
	d=0.4	0.204	0.26	0.411	0.182	0.228	0.335	0.118	0.135	0.226	0.042	0.058	0.127	0.04	0.053	0.125	0.033	0.044	0.107
	d=0.6	0.211	0.266	0.419	0.189	0.233	0.341	0.122	0.140	0.233	0.047	0.062	0.135	0.044	0.059	0.131	0.034	0.047	0.108
	d=0.8	0.219	0.271	0.429	0.195	0.239	0.353	0.125	0.142	0.241	0.049	0.067	0.137	0.045	0.06	0.132	0.036	0.050	0.110
84 Items	d=0	0.182	0.243	0.372	0.163	0.206	0.31	0.096	0.117	0.204	0.022	0.037	0.114	0.02	0.037	0.109	0.021	0.035	0.086
	d=0.2	0.187	0.248	0.379	0.168	0.213	0.318	0.101	0.122	0.213	0.026	0.04	0.118	0.024	0.041	0.113	0.023	0.036	0.091
	d=0.4	0.189	0.254	0.389	0.176	0.221	0.327	0.107	0.130	0.218	0.029	0.044	0.121	0.027	0.044	0.116	0.024	0.039	0.092
	d=0.6	0.195	0.261	0.396	0.183	0.230	0.339	0.115	0.138	0.224	0.033	0.047	0.125	0.03	0.047	0.120	0.026	0.041	0.097
	d=0.8	0.204	0.269	0.405	0.191	0.236	0.345	0.121	0.146	0.230	0.036	0.049	0.128	0.032	0.049	0.124	0.027	0.042	0.099
126 Items	d=0	0.168	0.235	0.357	0.152	0.195	0.298	0.084	0.106	0.192	0.015	0.022	0.098	0.015	0.021	0.077	0.013	0.02	0.065
	d=0.2	0.173	0.239	0.368	0.156	0.201	0.307	0.087	0.108	0.197	0.017	0.025	0.105	0.016	0.025	0.083	0.015	0.021	0.067
	d=0.4	0.179	0.244	0.378	0.158	0.208	0.317	0.088	0.112	0.204	0.022	0.029	0.109	0.023	0.031	0.086	0.016	0.023	0.069
	d=0.6	0.183	0.250	0.389	0.170	0.212	0.323	0.093	0.116	0.21	0.027	0.033	0.115	0.029	0.034	0.091	0.018	0.025	0.069
	d=0.8	0.19	0.259	0.393	0.174	0.219	0.329	0.095	0.122	0.217	0.031	0.038	0.119	0.029	0.037	0.097	0.018	0.026	0.072

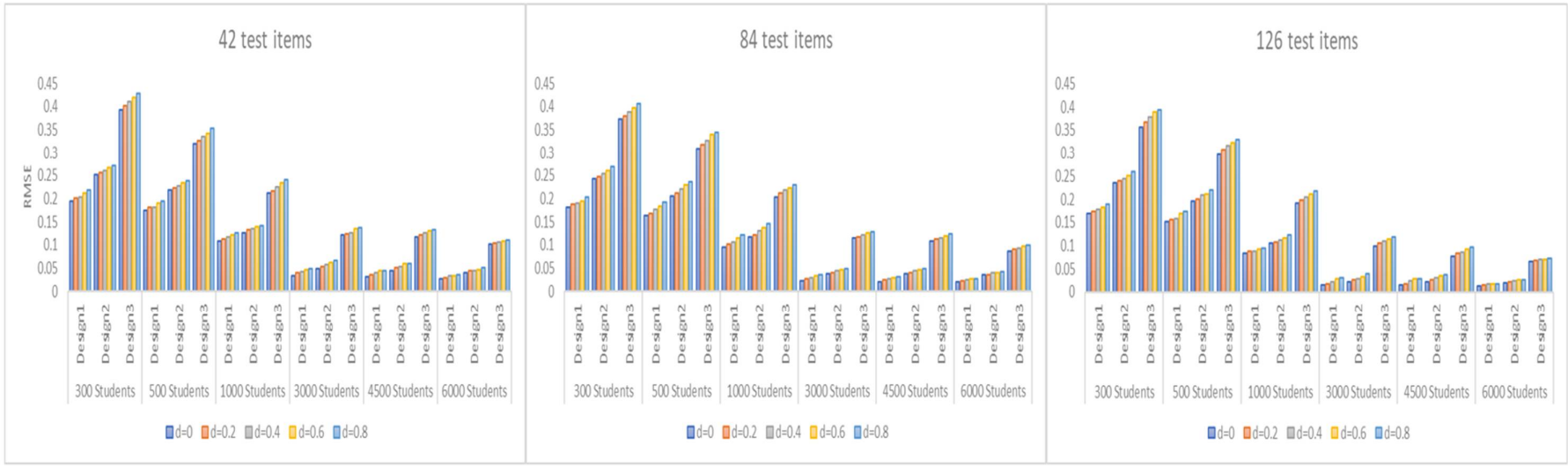
Note. "Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all the match conditions investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

Table 6.11. *Bias of recovered mean item difficulty across various levels of item-person match, sample size and test length*

	Match	300 Students			500 Students			1000 Students			3000 Students			4500 Students			6000 Students		
		Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3	Design1	Design2	Design3
42 Items	d=0	0.007	0.012	0.021	0.005	0.008	0.013	0.004	0.007	0.011	0.002	0.004	0.006	0.002	0.004	0.006	0.002	0.004	0.006
	d=0.2	0.008	0.012	0.022	0.005	0.009	0.014	0.004	0.007	0.012	0.002	0.004	0.006	0.002	0.004	0.007	0.002	0.004	0.006
	d=0.4	0.010	0.013	0.024	0.006	0.009	0.015	0.004	0.008	0.012	0.003	0.005	0.007	0.002	0.005	0.007	0.002	0.004	0.007
	d=0.6	0.010	0.013	0.024	0.007	0.01	0.015	0.005	0.008	0.013	0.003	0.005	0.008	0.003	0.005	0.008	0.002	0.005	0.007
	d=0.8	0.011	0.014	0.025	0.007	0.011	0.017	0.005	0.009	0.014	0.004	0.006	0.009	0.003	0.006	0.009	0.003	0.005	0.008
84 Items	d=0	0.004	0.007	0.015	0.002	0.005	0.009	0.001	0.003	0.006	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.004	0.007	0.016	0.002	0.005	0.010	0.001	0.003	0.006	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.005	0.008	0.017	0.002	0.006	0.011	0.001	0.003	0.007	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.005	0.009	0.017	0.003	0.006	0.011	0.001	0.004	0.007	0.000	0.000	0.003	0.000	0.000	0.001	0.000	0.000	0.000
	d=0.8	0.006	0.010	0.018	0.003	0.007	0.012	0.002	0.004	0.008	0.000	0.001	0.003	0.000	0.000	0.001	0.000	0.000	0.000
126 Items	d=0	0.001	0.002	0.007	0.000	0.001	0.004	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.2	0.001	0.002	0.007	0.000	0.001	0.004	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.4	0.002	0.003	0.008	0.000	0.002	0.005	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.6	0.002	0.004	0.009	0.001	0.002	0.005	0.000	0.000	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	d=0.8	0.003	0.004	0.010	0.001	0.003	0.006	0.000	0.001	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000

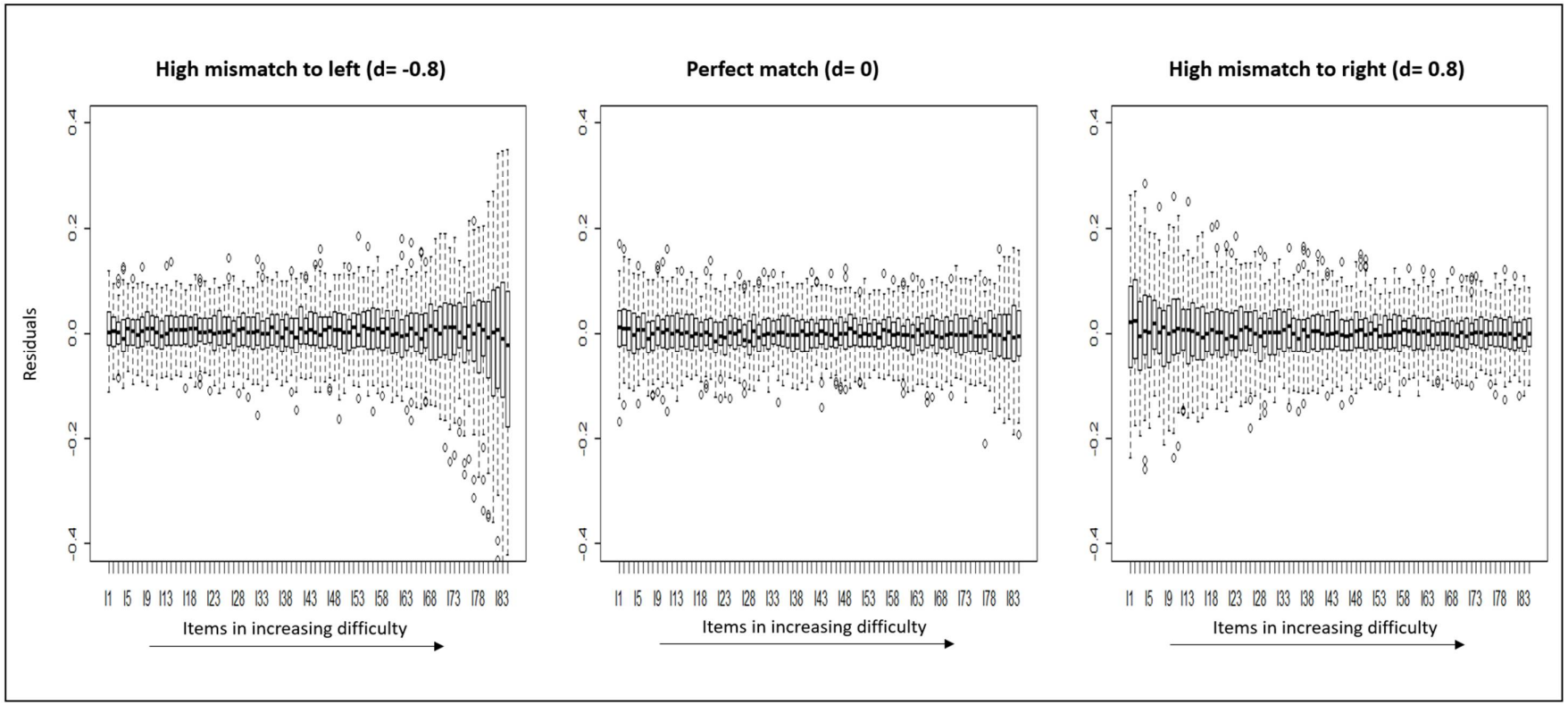
Note. "Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all the match conditions investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

Figure 6.5. *RMSE for the recovery of the mean item difficulty across different item-person match conditions, sample sizes and test lengths*



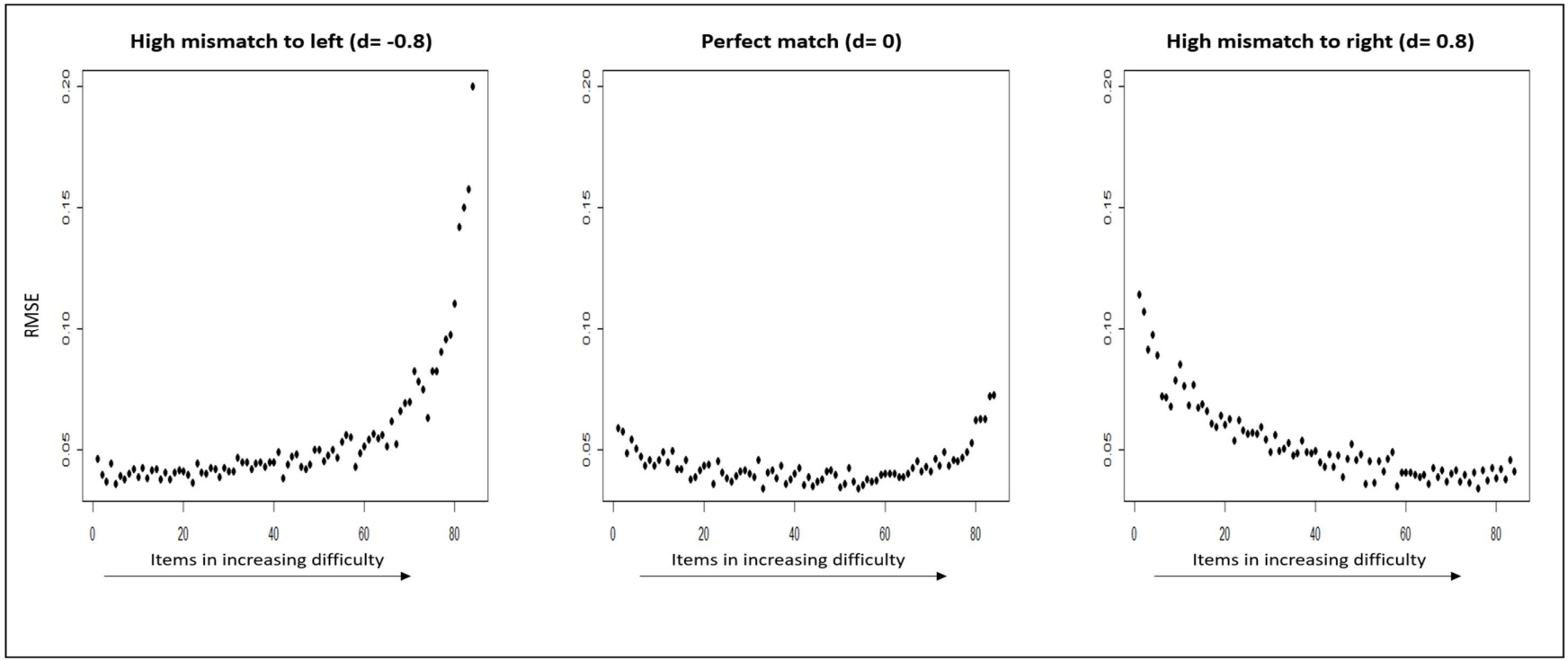
Note. The bar plots are in groups of 3's for each sample size (i.e., number of students). For instance, in the first panel which is a bar plot representing the case for a test length of 42 items, the first three groups are results for a sample size of 300 students. Further, each of these three groups represents results for one multi-matrix design. The multi-matrix designs are labelled D1, D2 and D3, with the designs becoming sparser moving from D1 to D3 (Multi-matrix Design D1 contains 57% missing data, while designs D2 and D3 contain 71% and 86% missing data respectively). Also, "Distribution Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). For the conditions d=0.2, d=0.4, d=0.6 and d=0.8, the mean for the distribution of person abilities is 0.2, 0.4, 0.6 and 0.8 respectively. In all distribution match conditions, the mean item difficulty is fixed at 0.

Figure 6.6. *Residuals between true item difficulty and the estimated item difficulty for individual items across 1000 simulations*



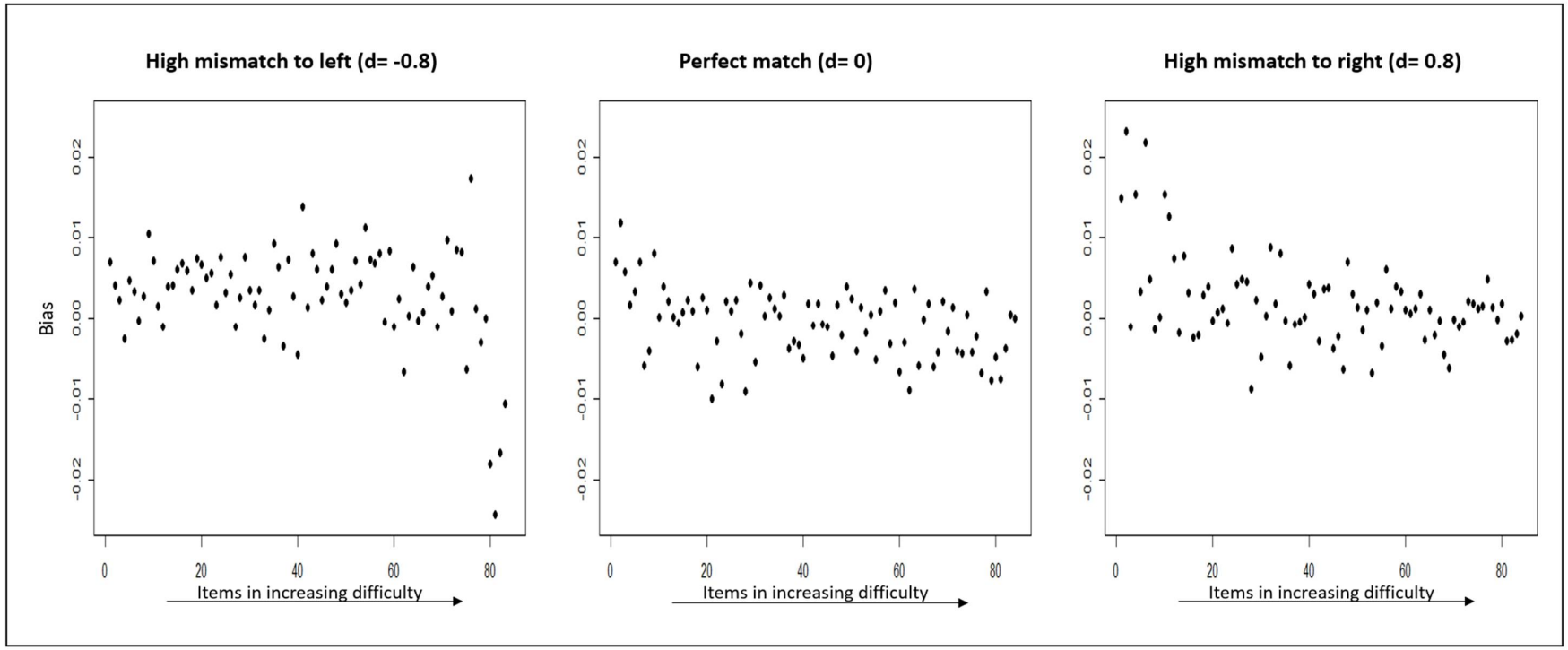
Note. The above results are for a test length of 84, sample size of 3000 students, and for the sparsest multi-matrix design (having 86% missing data). In all three cases (d=0.8, d=0, d=-0.8), the distribution of item difficulties has a mean of 0. However, in in Panel 1 (d= 0.8), the distribution of person abilities has a mean of 0.8. Similarly, the distribution of person abilities in Panel 2 (d=0) and Panel 3 (d=-0.8), have means of 0 and -0.8 respectively. Thus, in Panel 2 (d=0), the distribution of person abilities and the distribution of item difficulties are perfectly matched. Also, the test items are arranged in increasing order of difficulty with Item 1 being the easiest item, and Item 84 being the most difficult item.

Figure 6.7. *RMSE for the recovery of the item difficulty for individual items across 1000 simulations (N=3000 students)*



Note. In the three panels above (d=0.8, d=0 and d=-0.8), the distribution of item difficulties is kept fixed (with the mean item difficulty being 0); while the distribution of person abilities is shifted to the right in Panel 1, d=0.8 (so, that the mean person ability is 0.8); and in Panel 3, d=-0.8, the distribution of person abilities is shifted to the left (so that the mean person ability is -0.8). In Panel 2 (d=0), the distribution of person abilities and the distribution of item difficulties are perfectly matched, both distributions having a mean of 0. Further, the results are for the case where the test length is 84, sample size of 3000 students, and with the sparsest multi-matrix design (i.e., the multi-matrix design having 86% missing data). Also, the 84 test items are arranged in increasing order of difficulty. So, Item 1 is the easiest item, while item 84 is the most difficult item.

Figure 6.8. *Bias for the recovery of the item difficulty for individual items across 1000 simulations (N=3000 students)*



Note. In the three panels above (d=0.8, d=0 and d=-0.8), the distribution of item difficulties is kept fixed (with the mean item difficulty being 0); while the distribution of person abilities is shifted to the right in Panel 1, d=0.8 (so, that the mean person ability is 0.8); and in Panel 3, d=-0.8, the distribution of person abilities is shifted to the left (so that the mean person ability is -0.8). In Panel 2 (d=0), the distribution of person abilities and the distribution of item difficulties are perfectly matched, both distributions having a mean of 0. Further, the results are for the case where the test length is 84, sample size of 3000 students, and with the sparsest multi-matrix design (i.e., the multi-matrix design having 86% missing data). Also, the 84 test items are arranged in increasing order of difficulty. So, Item 1 is the easiest item, while item 84 is the most difficult item.

Table 6.12. *ANOVA results with log(RMSE) of recovered mean item difficulty as criterion*

Source	df	Mean square	F value	η^2
Sample Size	5	5.905	5446.27	.70
Test Length	2	0.482	444.72	.02
Matrix Sparseness	2	4.719	4352.52	.22
Distribution Match	4	0.074	68.37	.01
Sample Size X Test Length	10	0.040	36.97	.01
Sample Size X Matrix Sparseness	10	0.109	100.89	.03
Test Length X Matrix Sparseness	4	0.021	19.05	.00
Residuals	232	0.001		

Note. $p < .001$ for all cases.

Chapter 7 General discussion

In the previous chapter, a detailed report was presented on the results of this doctoral project. The current chapter consists of a summary on studies carried out; a discussion of the findings observed; implications of these findings to educational practice; as well as study limitations and recommendations for further research. The chapter's purpose will be to provide an in-depth discourse on how factors investigated in the project (i.e. test length, sample size, sub-group populations, and item-person match) relate to efficiency with which person and item parameters are recovered. This will be followed by a discussion of challenges encountered in carrying out the studies; as well as consequences or implications of findings to test developers, measurement experts and education policy makers. At the end of chapter, a general summary will be given to capture the substance and scope covered by the entire project, as well as directions for future research and practice.

7.1 Summary of studies

This doctoral project aimed at investigating how the efficiency with which various sparse multi-matrix booklet designs recovered item and population parameters. In order to do this, factors such as the number of students, number of items, item-person match; as well as, sub-groups within the population, were examined to learn how these relate to the efficiency with which person and item parameters were recovered. The person parameters investigated were the mean and variance of the distribution of person abilities; while the item parameter investigated was the item location parameter (i.e., the item difficulty). It is important to note that in large-scale assessments, emphasis is not on the performance of individual students, but

how groups of students perform. That is why in this project individual student ability was not considered, but rather the mean and variance of the distribution of person ability of populations or sub-populations of students. Further, Balanced Incomplete Block-7 booklet Designs like those in von Davier et al. (2009) and Gonzalez and Rutkowski (2010) were used. These designs possess several beneficial characteristics (like every item-pair combination occurring at least once, and an equal number of times) and variants of these designs used in several large-scale assessments like PISA (Gonzalez & Rutkowski, 2010).

Thus, this doctoral project answered the following key research questions:

1. How efficiently are item and person parameters recovered at the global population level in the various sparse matrix booklet designs?
2. How is test length and sample size related to the efficiency and precision with which person and item parameters are recovered in the various sparse matrix booklet designs?
3. How efficiently are performance differences between policy relevant population subgroups recovered when using the various multi-matrix booklet designs (across various conditions investigated)?
4. Considering test length, how is efficiency of recovered person and item parameters influenced by the match between item and person ability distributions in various sparse matrix booklet designs?
5. Considering sample size, how is efficiency of recovered person and item parameters influenced by the match between item and person ability distributions in various sparse matrix booklet designs?

These research questions were answered under two large studies, in which Study 1 answered the first three research questions, while Study 2 answered the last two research questions. In Study 1, real data (VERA-8 2015 mathematics assessment for Berlin and Brandenburg) and simulated data were used, while for Study 2 simulated data was used. The data for Study 2 were simulated to have properties like the PISA 2012 Mathematics data for Germany.

Further, person an item parameters were scaled using the mixed coefficient multinomial logit model (MCMLM; Adams & Wu, 2007) which is a generalized multidimensional Rasch model; while the RMSE (root mean squared error) and bias of the recovered person and item parameters were used to examine parameter recovery efficiency. This was done using 1000 replications in each experimental condition to ensure stable and reliable results.

In general, the results showed that:

- At the global population level (for the VERA-8 dataset), the mean and the variance of the distribution of person ability were recovered accurately and without bias (RMSE ≤ .04). However, the mean of the distribution of item difficulties was inaccurately recovered especially when using the sparsest multi-matrix booklet design.
- Test length and sample size were consistently related to the precision with which the various matrix designs recover person and item parameters of interest. However, increasing the sample size beyond 3000 students led to insignificant gains in parameter recovery precision.
- Performance differences between population subgroups were recovered accurately and without bias, across all matrix booklet designs and conditions when sample size was at least 3000 students.
- The degree of match between the distribution of person abilities and the distribution of item difficulties affected parameter recovery precision especially when the sample size was less than 1000 students. However, this effect reduced greatly with increasing sample size. Thus, after a sample size of more than 3000 students, the effect became almost negligible.

7.2 Discussion of findings

A detailed discussion of findings of this project are presented under three broad headings. The first heading will discuss how test length (i.e., number of items) and sample size (i.e., number of students) are related to the efficiency and precision with which item and person

parameters are recovered in the various sparse booklet designs. The second heading will proceed to elaborate on efficiency of item and person parameter recovery at the group or sub-population level; while the last section will tackle how the match between the distribution of person abilities and the distribution of item difficulties impacts parameter recovery efficiency (taking into consideration test length and sample size).

7.2.1 Test length, sample size, and parameter recovery efficiency in sparse matrix booklet designs

To gain the utmost benefits of IRT, it is important to ensure accurate estimation of IRT model parameters (He & Wheadon, 2013; Kieftenbeld & Natesan, 2012). Importantly, a myriad of factors and conditions—including test length and sample size—can influence the precision or accuracy with which item and person parameters are recovered (Hambleton, 1989; Wollack et al., 2006). Further, numerous studies have investigated how test length and sample size impact parameter recovery in an IRT context (e.g. see De Mars, 2003; He & Wheadon, 2016; Sahin & Anil, 2016; Swaminathan et al., 2003; Tay, Huang & Vermunt, 2016; Wang & Chen, 2005). The general conclusion from these studies is that short tests with few students result in less precise item and person parameter estimates, in contrast to long tests with large samples which produce more precise parameter estimates.

Thus, a resulting—and pertinent—question from the above is, investigating appropriate sample size and test length requirements, for obtaining acceptable levels of precision for estimated person and item parameters in an IRT context. Several studies have been carried out to investigate this problem and there have not been any clear-cut recommendations on the number of test participants or number of items required for person or item parameters to be recovered accurately. This is more the case because researchers use different benchmarks to evaluate what they consider as an accurate level of recovery and the requirement for such accuracy differs with respect to the IRT models used. Table 7.1 below gives a summary of some studies that investigated this problem and the minimum sample size that was recommended. It is important to note that the studies all involve the use of complete datasets without any application of multi-matrix booklet designs.

Table 7.1. *Sample size recommendations from some IRT studies on parameter recovery*

Model	Study	Recommended minimum sample size for accurate parameter recovery
Dichotomous 1 PL Model	Thiessen & Wainer (1982)	N ≥ 500
	Goldman & Ruju (1986)	N ≥ 250
	Gruyer & Thompson (2011)	N ≥ 300
2 PL Model	Hulin, Luissak & Drasgrow (1982)	N ≥ 500 and 30 items
	Lim & Drasgrow (1990)	N ≥ 500 and 20 items
	Harwel & Janosky (1991)	N ≥ 500 and 25 items
	Stone (1992)	N ≥ 500
	Yoes (1995)	N ≥ 300 and 75 items
	Wiess & Minden (2012)	N ≥ 300
3 PL Model	Lord (1982)	N ≥ 1000 and 50 items
	Swaminathan & Gifford (1979)	N ≥ 1000
	Yen (1987)	N ≥ 1000
	Tang, Way & Carey (1993)	N ≥ 1000
	Yoes (1995)	N ≥ 1000
	Patsula & Gessaroli (1995)	N ≥ 1000

Thus, interesting questions that result from the above are examining sample size requirements for acceptable levels of parameter recovery precision in different multi-matrix booklet designs; as well as, further examining how test length and sample size relate to parameter recovery efficiency in these designs. Unfortunately, the knowledge base on multi-matrix designs and parameter recovery is still very limited, with "much of the discussions around multi-matrix sampling having been relegated to the pages of technical manuals" (Rutkowski, Gonzalez, von Davier, & Zhou, 2014, p.76). In a slightly related study, Gonzalez and Rutkowski (2010) performed a simulation study using balanced incomplete block booklet designs; and examined the degree to which item and population parameters were recovered in relation to matrix sparseness and sample size. They found that when the booklets had fewer items, person abilities became less accurate and as the number of test participants increased, the precision with which item difficulties were recovered increased (though the gain in precision was greater for difficult items as compared to the easier items). However, this important study used completely simulated data. This kind of data often fits perfectly to underlying IRT models used in the study (unlike in real empirical data), thus raising concerns about generalizability of the results in real data. Further Gonzalez and Rutkowski (2010) used only one sample size and test length.

To fill this gap in the literature, the first study of this dissertation project carried out a study like that of Gonzalez and Rutkowski (2010) with the difference that it uses empirical assessment data; and, incorporates several levels of test length and sample sizes in the design of the study. Using different test lengths and sample sizes makes it possible not only to examine parameter recovery efficiency in the various matrix booklet designs; but, also get sample size requirements for acceptable levels of precision for the recovery of population and item parameters of interest.

As expected, and in line with previous literature on test length, sample size, and parameter recovery, the study results showed that as sample size and test length increased, the precision with which population and item parameters were recovered improved. Importantly, sample size accounted for the greatest amount of variance in the RMSE of all recovered person and item parameters investigated (explaining more than 50% of the total variance). However,

increasing the sample size beyond 3000 led to very little gains in parameter recovery precision. Generally, the mean of the distribution of person ability was recovered accurately[17] by all multi-matrix booklet designs when the sample size was at least 1000 test participants. However, for recovery of the variance of the distribution of person ability, a sample size of at least 3000 subjects was required for accurate parameter recovery when using the sparsest multi-matrix booklet design. Similarly, to recover the mean item difficulty accurately, a minimum sample size of 3000 test participants were required, irrespective of the multi-matrix booklet design used.

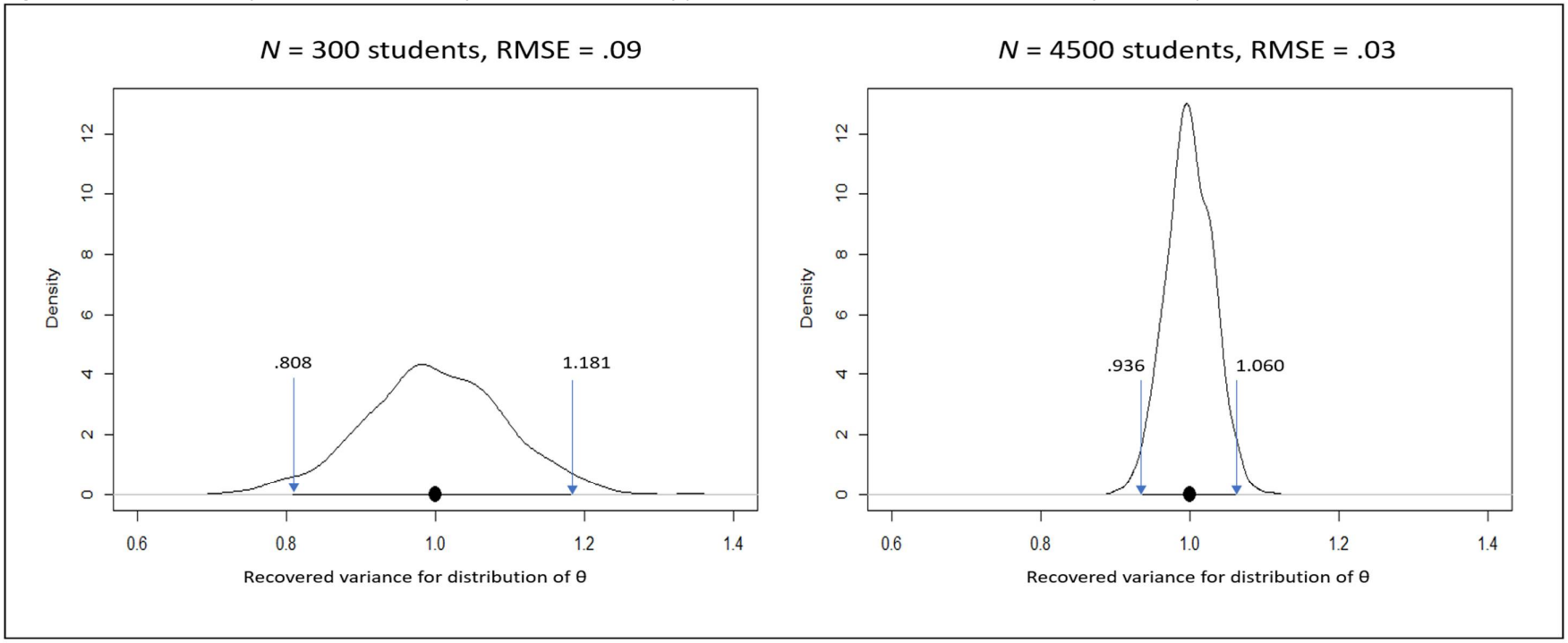
On the other hand, although parameter recovery accuracy improved with increasing test length, the gain in precision as a result of the increase in test length was small. For instance, ANOVA analysis of the results showed that test length accounted for only 1% of the variance in the log(RMSE) of the recovered mean for the distribution of person abilities. Similarly, it accounted for 8% and 2% respectively for the variance in log(RMSE) for the recovered variance of the distribution of person abilities and mean item difficulty. It is important to note that test on its own cannot be used as a benchmark for describing parameter recovery accuracy but makes sense only when described for a given sample size. Thus, as an example, with at least 42 items and 3000 test participants, the variance of the distribution of person abilities and the mean item difficulty can be recovered accurately, when using any of the multi-matrix booklet designs. Further, to visualize how the RMSE and the accuracy of a given parameter are related, Figure 7.1 below illustrates two cases in which in the first case, the variance of the distribution of person abilities is recovered accurately across 1000 simulations, while in the second case, this person parameter is not recovered accurately.

It is interesting to note that the results from the first study of this dissertation were like those from the second study, although both studies used different datasets. The results for test length, sample size and parameter recovery for the first study, can be considered a subset of the results from the second study—for all conditions where a perfect match exists between the distribution of person abilities and the distribution of item difficulties. The gain in

[17] An item or population parameter was considered as being accurately recovered, when the RMSE of the recovered parameter was $\leq .04$

parameter recovery precision with increasing sample size results from the increase in the amount of information available for the estimation of population and item parameters. However, continuously increasing the sample size will lead to a point where the parameter estimates become so accurate that further sample size increases result in very little gain in parameter recovery precision. This threshold was found to be a sample size of about 3000 subjects when using the multi-matrix designs and conditions examined in this project.

Figure 7.1. *Distribution of recovered variance (for the distribution of person abilities) across 1000 simulations for two experimental conditions*



Note. The results are for the case where multi-matrix Design1 is used (i.e., the Design with 57% missing data) and when there is perfect overlap between the distribution of person abilities and the distribution of item difficulties. The true variance of the distribution of person abilities was 1. In Panel 1 (where *N*=300 students), the variance is not accurately recovered, with the 95% confidence interval of the recovered variance (for the distribution of person abilities) lying in the range [.808, 1.181]. On the other hand, in Panel 2 (where N=4500 students), the variance (for the distribution of person abilities) is recovered accurately, with the 95% confidence interval for this person parameter lying in the range [.936, 1.060].

7.2.2 Group level parameter recovery

While performing large-scale assessments, one crucial objective is to examine performance related disparities that exist between relevant subgroups within populations. Such important subgroups within the population could be classified based on relevant educational, geo-political or demographic variables such as gender, ethnicity or socio-economic status. Taking the United States as an example, the Every Student Act (ESSA) of 2015 (Public Law 114-15) imposes state-wide accountability, wherein educational outcomes of students from major ethnic and racial groups, economically disadvantaged students, English learners, and disabled students are systematically reported (Seastrom, 2017). Further, the "No Child Left Behind Act" (NCLB) demands that schools be held accountable for the performance of the school, as well as for designated subgroups, starting with the 2002-2003 academic year (Simpson, Gong & Marion, 2006).

One long standing criticism of large-scale assessments is the existence of substantial and persistent score disparities between test takers with minority and nonminority backgrounds (Bronner, 1997; Jencks & Philips, 1998; Sacks, 1997). The troublesome issue of persistent score differences by socioeconomic status (SES) and among racial or ethnic groups have led to some charges of test bias (Camara & Schmidt, 1999).

It is however not surprising that, grouping individuals in different ways that are associated to differences in their access to educational opportunities, could result in group members scoring differently. It is thus noteworthy that differences in test mean scores or on other measures, are not necessarily a measure of bias—the more crucial concern in large-scale assessments being whether differential predictive ability exists among the concerned groups (Camara & Schmidt, 1999).

Given that performance differences can exist between subpopulations or groups, an important measurement or policy objective is to ensure that these performance differences between groups or subpopulations are well estimated and uncovered whenever they exist. To achieve this goal, several multi-group IRT models have been developed (e.g., Adams, Wang & Kang, 1997; Béguin & Glas, 2001; Steenkamp, De Jong, & Baumgartner, 2010; Mislevy, 1983;

Padilla, Azevedo & Lachos, 2018). However, it is possible that the amount of missing data in a multi-matrix booklet design (i.e., matrix sparseness) impacts the precision with which item and population parameters are recovered. In a related simulation study, Gonzalez and Rutkowski (2010) examined the efficiency with which person and item parameters were recovered at the group level when using various sparse multi-matrix booklet designs. Using EAP scores their results showed that as the multi-matrix booklet designs became sparser, person ability estimates became less reliable, and group differences in the population became underestimated, when these existed.

In this PhD project, the above question examined by Gonzalez and Rutkowski (2010) was further examined. In this case, plausible values were used instead of EAP scores (which are point estimates) for estimating person abilities; also, unlike using completely simulated data, real assessment data were used in simulating the test data—hence, the resulting simulated data had properties of the real assessment dataset. However, the results of the study were different from those of Gonzalez and Rutkowski (2010) in that group differences in the population were recovered accurately and without bias when the sample size was at least 3000 test participants. Particularly, group differences in the mean of the distribution of person abilities were recovered accurately with a minimum sample size of 1000 test participants ($0.015 \leq RMSE \leq 0.022$ and $0.000 \leq Bias \leq 0.002$). However, for accurate recovery of group differences in the variance of the distribution of person abilities, a minimum sample size of 3000 test takers was required ($0.019 \leq RMSE \leq 0.044$ and $0.000 \leq Bias \leq 0.007$).

Further, huge performance differences slightly affected the precision of recovered person parameters especially when the sample size was less than 3,000 students. For instance, when the sample size was 300 students and for the sparsest multi-matrix design, the increase in the RMSE of the group difference in the mean of the distribution of person abilities was 0.012, when the two groups had no performance differences (d=0 condition) and when the two groups had the greatest amount of performance differences (d=2 condition). However, when the sample size was now 10,000 students, the RMSE only increased by a negligible amount of 0.002 when those two same conditions were considered (i.e., the case of no performance

differences between the two groups, d=0; and the case with the greatest performance differences between the two groups, d=2).

The above results thus support the use of plausible values in conjunction with multi-matrix booklet designs in estimating student performance at group or subpopulation levels in large-scale assessments (where the sample size is usually greater than 3,000 students). This procedure produces accurate and unbiased recovery of mean student performance in groups or subpopulations; unlike when point estimates (e.g., EAP scores) are used in conjunction with the booklet designs to estimate student performance at group or sub-population levels. As noted by Mislevy et. al. (1992), a challenge in large scale educational assessments is the fact that distributions of point estimates—that would be preferred for making inferences about individuals—can depart substantially from distributions of underlying latent variables investigated.

7.2.3 Item-person match and parameter recovery

In order to measure examinee performance most effectively, not-so-easy or not-so-difficult items need to be administered to test takers (Lord, 1980). This implies, ideally, in a population of students with varying abilities, item sets or test booklets with varying difficulties are required for efficient measurement of individual student abilities (Weiss, 1982). However, although students are administered different sets of items—which could be from easy, moderate or even difficult test booklets—the final test scores need be directly comparable (Berger et al., 2019). This goal is however easily achieved using one of several test equating strategies (Kolen & Brennan, 2014).

Further, taking for instance the one-parameter logistic model, the quality of parameter estimates ought to degrade to the extent that person and item location parameters fail to match to one another, since individual items are maximally informative for person parameter estimation—and individual persons maximally informative for item parameter estimation—when an item and person lie at the same point on the latent trait continuum (Svetina et al., 2013). In a large-scale assessment scenario, such person-item mismatch can occur when easy

items are administered to a population of high ability students or difficult items administered to a population of low ability students.

Unfortunately, "considerably less research has been conducted investigating IRT methods where a mismatch between item and person parameter distributions exist" Svetina et al. (2013, p.336). As might be expected, measurement efficiency could be affected by factors such as boredom or lack of motivation, which is possible when test items fail to fit student abilities. This could consequently impact on the precision with which population and item parameters are recovered. For instance, Asseburg & Frey (2013) investigated the relationship between ability-difficulty fit (i.e., match between person ability and item difficulty) and effort or boredom. They used 9,452 ninth graders in Germany (PISA 2006) who took a mathematics test and responded to a questionnaire on test-taking effort (motivation) and boredom/daydreaming (emotion). Their results showed that ability-difficulty fit was positively linear-related with effort and boredom/daydreaming. In a more related study, Svetina et al. (2013) performed a simulation study to investigate recovery of item and person parameters of the one-parameter logistic model for short tests administered to small samples. They manipulated the match between the distribution of person abilities and the distribution of item difficulties, as well as test length, sample size, and item discrimination. Their results showed that match between the distributions of person abilities and the distribution of item difficulties had a modest effect on parameter recovery; and accuracy in parameter estimation decreased as sample size and test length decreased.

Thus, to fill the research gap on a dearth of literature investigating IRT methods where a mismatch between item and person parameter distributions exists, the second study of this PhD dissertation focused on examining item-person match and parameter recovery in sparse multiple matrix booklet designs. Unlike in Svetina et al. (2013) where completely simulated and complete datasets were used, our study tackled the question from a different perspective using data simulated from an empirical dataset, and further containing missing data due to treatment with multiple matrix sampling. Completely simulated data usually fit underlying IRT models more perfectly, and small effects discovered using complete data could become larger when using incomplete data with missingness.

Importantly, even though the results showed that item-person match impacted parameter recovery precision, with greater item-person match resulting in more precise parameter estimates, this effect was very small. For instance, the ANOVA results showed that item-person match accounted for less than 1% of the total variance in log(RMSE) of recovered mean for distribution of person abilities (other sources of variation in the model being test length, sample size, sparseness in booklet design, and the interaction between sample size and booklet design sparseness). Similarly, for recovery of mean item difficulty and variance of distribution of person ability, item-person match only explained 1% and 5% of the total variance in the log(RMSE) respectively.

Further, the ANOVA results also showed a very small but significant interaction between sample size (i.e., number of test participants) and the degree of sparseness in the booklet design. This interaction accounted for up to 3% of the variance in log(RMSE) of the recovered mean item difficulty, with item-person match having a stronger effect when the test length was less than 1000 test participants. For example, considering the sparsest multi-matrix design (Design3) and 42-itemed test, the RMSE of the recovered mean item difficulty increased by 0.037 logits from the perfect match case (d=0) to the most mismatched case (d=0.8), for a sample size of 300 test takers. However, when the sample size was 6000 test takers, this increase in RMSE was only 0.009 logits (i.e., between the perfectly matched case and the most mismatched case, under the same conditions).

On the other hand, it was also interesting to look at how item difficulties were recovered at individual item levels. The results showed that depending on the direction of item-person match, various extremes of the item difficulty continuum (i.e., either very easy or very difficult items) were affected differently. For instance, for the distribution match condition, d=0.8, where the distribution of person abilities had a mean of 0.8, while the mean of the distribution of item difficulties remained fixed at 0; the results showed that the more difficult items were recovered more accurately than the easier items. However, for the distribution match condition, d=0, where both distributions of item difficulties and person abilities were perfectly aligned (with both distributions having a mean of 0), the results showed that the item difficulties for all items were recovered with almost the same degree of accuracy. This

could be explained by the fact that shifting the distribution of person abilities to the right (e.g. case where the mean of the distribution of person abilities was 0.8), with the distribution of item difficulties remaining fixed (with a mean of 0), results in less information being available for estimating difficulties for the very easy test items.

It is noteworthy that, the fundamental idea behind targeted testing designs like computerized adaptive testing is ensuring a good match between the distribution of item difficulties and the distribution of person abilities. Such targeted testing designs increase measurement efficiency since, test items are not too easy to make the test takers bored; nor too difficult to reduce their motivation for taking the test (especially in the case of low stakes assessments like in most large-scale assessments). However, it is suggested that some caution be exercised when comparing item difficulties from worst performing and best performing countries in a large-scale assessment, since the precision with which very easy and very difficult items are recovered could be slightly different in both countries. This could pose a challenge when calculating the proportion of students belonging to a certain proficiency level, judging the cognitive demands of items, or doing standard setting procedures based on the empirical item difficulty.

7.3 Implications of research findings

This section relates how findings of this dissertation are applicable in practice. These implications are given under two broad headings a) the implications to test developers and measurement experts and b) the implications to education policy makers.

7.3.1 Implications for test developers and measurement experts

Many empirical investigations suggest that test length and sample size influence efficiency with which item and person parameters are recovered in IRT (e.g., see Akour & Al-Omari, 2013; Chuah, Drasgow & Luecht, 2006; DeMars, 2003; He & Wheadon, 2013; Sahin & Anil, 2016; Stone & Yumuto, 2004; Swaminathan et al., 2003; Toland, 2008). An important challenge to test developers and measurement experts is, thus, finding sample size

requirements for accurate estimation of given population or item parameters when applying IRT models. Further, even if parameter recovery *"accuracy*" is not the main concern, it is often still interesting to investigate parameter recovery *"efficiency"* (i.e., how to recover item and population parameters accurately, while at the same time minimizing the amount of resources used to achieve this).

From the results of this dissertation, it can be recommended that, to accurately recover person parameters (i.e., the mean and variance of the distribution of person abilities), a sample size of more than 1000 test participants and 100 test items be used when applying multi-matrix booklet designs like those in this PhD project. However, for accurate recovery of the item parameter (i.e., the mean item difficulty), it is recommended that a minimum sample size of 3000 test participants and 100 items be used, when applying a booklet design like Design1 (with 57% missing data) or Design2 (with 71% missing data). Importantly, as the sample size increases, its association with the RMSE of recovered item or population parameters diminishes.

On the other hand, test length, sparseness in the booklet design and distribution match are also found to be associated with the accuracy with which item and person parameters are recovered. Longer tests result in more precise parameter estimates. For instance, doubling the test length from 42 to 84 test items results in a general reduction of 25% in the RMSE of the recovered mean for the distribution of person abilities. However, when test length is further doubled from 84 to 126 test items, the RMSE (for the mean of the distribution of person abilities) only reduces by about 11%. Thus, as test length increases, the associated gain in parameter recovery precision becomes less.

Similarly, as expected, matrix sparseness (i.e., the amount of missing data in a booklet design) is also found to relate to the precision with which person and item parameters are recovered, with a sparser booklet design resulting in less precise population and item parameter estimates. Thus, a reduction of about 30% in matrix sparseness (i.e., from the least sparse to the sparsest of the booklet designs considered) results in a mean increase of 0.14 in the RMSE of the recovered mean for the distribution of item difficulties. However, when considering recovery of the mean and variance for the distribution of person abilities (for the

same conditions as above), this mean reduction in RMSE only becomes 0.003 and 0.06 respectively. Thus, matrix sparseness has the greatest impact on the recovery precision of the mean for the distribution of item difficulties.

On the other hand, although the match between the distribution of person abilities and the distribution of item difficulties is found to be associated with the precision with which item and population parameters are recovered, the strength of this association is very weak. For instance, a reduction of about 30% in the match between the two distributions (i.e., comparing the case where the two distributions match perfectly with the most mismatched case) results in a negligible mean increase of about 0.001, in the RMSE for the recovered mean for the distribution of person abilities. Similarly, when considering recovery of the mean item difficulty and the same conditions, the mean increase (in RMSE) was just about 0.02. Therefore, in designing large scale assessments, item-person match should not be considered a major challenge to parameter recovery accuracy, especially when the sample size is above 3000 participants, since its effect on parameter estimates has been found to be negligible.

7.3.2 Implications for policy makers, politicians and other stakeholders in Education

Large scale educational assessments provide information on a system's educational outcomes and—if questionnaires are administered to get background information from students, teachers, parents, and/or schools—the associated factors, which can thus help policy makers and other stakeholders in the education system in making policy and resourcing decisions for improvement (Anderson, Chiu & Yore, 2010; Benavot & Tanner, 2007; Braun, Kanjee & Bettinger, 2006; Grek, 2009; Postlethwaite & Kellaghan, 2008). This perspective to education policymaking, based on evidence, including data from large-scale assessments has been adopted around the world (Lietz, Cresswell, Rust, & Adams, 2017); and has not only become the most frequently reported method used by politicians and policy makers, but now considered a global norm for educational governance (Wiseman, 2010). Further, evidence-based policy making involves measuring and ensuring quality, ensuring equity, and

accountability (Lietz et al., 2017; Wisemann, 2010). In order to provide indicators of equity, it is necessary to compare the performance of policy relevant sub-groups within populations—for instance, based on gender, socio-economic status or ethnic background (Lietz et al., 2017). Accurate measurement of performance differences between policy relevant population subgroups is thus a crucial objective in large-scale assessments.

Therefore, results from this dissertation are important to education policy makers since they offer further support that multi-matrix booklet designs can be used to accurately estimate student performance in subgroups within populations. By accurately estimating the performance of population subgroups (for instance, in terms of gender, socio-economic status or school type), performance gaps or differences existing between these subgroups can be exposed, thus providing evidence for education policy makers and politicians to create policies or legislation geared towards bridging such inequity.

Further, in large-scale assessments, policy makers are also interested in the distribution of person ability, including its mean and variance. This is because it can show the proportion of students that belong to certain proficiency levels, which in turn could guide policy making (for instance, the need to allocate more resources to help low ability students, or give prizes to very high achieving students). The results are thus relevant to policy makers, as they show that in a large scale assessment (with more than 1000 students and at least 100 items), the mean and variance of the distribution of person abilities is recovered accurately, when using multi-matrix designs like those investigated in this PhD project.

On the other hand, when conducting large-scale assessments or surveys, there is usually a need to balance topical breadth and depth with factors related to respondent behaviour, such as compliance, motivation, and concentration. On one hand, there is need to obtain the greatest amount of information on subjects in the sample, in order to ensure better modelling of complex human perceptions, attitudes and behaviour. On the other hand, economic restrictions, as well as psychological and motivational factors must guide construction of any test (Wolf, 2006). Lengthy tests can thus be problematic in that they require more resources to construct and administer; place a heavier burden on the respondents; and, this could result

in increased non-response, modified answering behaviour, and thus, greater measurement error.

The results of this dissertation yield further support for the use of multi-matrix booklet designs as a very efficient form of test abridgement in large scale assessments. Thus, by applying multi-matrix booklet designs, shorter tests can be developed, resulting in significantly reduced test construction and administration costs. Particularly, the results show that by using a multi-matrix booklet design like Design 3, test length can be reduced by 86 percent and yet accurate item and person parameters still recovered when using a sample of more than 1000 students and at least 100 test items). This can result in massive savings of both financial and material resources required for test construction and administration. Further, shorter tests are better for the respondents since they involve lesser respondent burden which in turn could further motivate them to complete the entire test. On the other hand, since shorter tests require lesser time to complete, school principals could be more willing to allow their students to partake in such assessments, since they will result in lesser disruption on the school's timetable.

Importantly, even though the results show that in large scale assessments item-person match does not pose a great challenge to the precision of population estimates (i.e., the mean and variance of the distribution of person abilities), there is need to be wary when considering item parameters. This is because a high mismatch between the distribution of person abilities and the distribution of item difficulties can result in less precise item location estimates for extreme items (i.e., very easy items or very difficult items). For instance, administering a test of average difficulty to a population of very high achieving students, could result in very easy items being less accurately estimated. On the other hand, in a population of low achieving students, the difficult items could be estimated less accurately.

7.4 Study limitations and recommendations for future research

Despite the merits of the results of this project (e.g., on the relationship between item-person match and parameter recovery efficiency), as expected with studies of this magnitude, some

limitations still exist. For instance, during person and item parameter estimation, item-position effects were not fully taken into consideration. Importantly, several studies show that in performing large scale assessments, a source of bias during parameter estimation are effects resulting from the position items are presented in a booklet (e.g., Albano, 2013; Debeer & Janssen, 2013; Hahne, 2008; Hohensinn, Kubinger, Reif, Holocher-Ertl, Khorramdel, & Frebort, 2008; Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011; Weirich, Hecht, & Böhme, 2014). Usually, the applied test design contains several test booklets with the same items presented at different test positions (Hohensinn et al., 2008). Students thus answer one of several booklets with the order in which items are presented in each booklet being different. This variation in item positions within booklets could potentially affect the probability of a correct response (Hecht, Weirich, Siegle, & Frey, 2015b).

This phenomenon—referred to as position effects—is interpretable from either a person or item perspective. From an item perspective, item parameters such as item difficulties are seen to depend on the item position (e.g., an item may be found to be more difficult if administered towards the end of a test). From a person perspective, an examinee's competence estimate may be seen to drop towards the end of a test causing estimated competencies to be greater at the beginning of the test than towards the end. This could be explained by the effects of fatigue, motivational aspects, or training effects (Hecht et al., 2015b). Examinee performance could decrease towards the end of the test because they become more exhausted and demotivated, or conversely, increase because they become more accustomed to the kind of test material being used.

Therefore, it could be interesting to conduct a similar study using a model that takes into consideration item position effects when estimating person and item parameters. However, it is important to note that using a balanced incomplete block design in this project helped to partly mitigate the problem caused item-position effects. As noted by Frey & Bernhardt (2012), position and carry-over effects are not removed when using a balanced incomplete block design but only averaged across positions. Generally, this is not a problem in most large-scale assessments, since emphasis is not on valid individual ability estimation, but on

ability estimation at group or subpopulation levels. However, some large-scale assessments give additional feedback to every individual test taker, and sometimes to individual classes or schools; in such cases, when several test booklets are used, for instance to limit cheating, any item-position effect could invalidate results obtained for any individual test taker (Hohensinn et al., 2008).

Another limitation encountered in the course of this project was that person parameter estimation problems were encountered when the sample size was less than 300 and when using the sparsest multi-matrix booklet design. This resulted because the MML estimator sometimes failed to converge. It could have been interesting to find a lower threshold (to serve as the sample size requirement) for accurate recovery of the mean person ability in the various booklet designs. On the other hand, there are so many possibilities as to how balanced incomplete block designs can be constructed. For instance, the balanced incomplete block design used in the PISA 2006 assessment contained 14 booklets and 4 item blocks (Frey & Bernhardt, 2006). However, this project examined only three kinds of balanced incomplete block designs. Other studies could thus be carried out using different forms and variants of balanced incomplete block designs to verify if similar results will be obtained.

In the second study of this project (on item-person match and parameter recovery efficiency), it was assumed that person abilities and item difficulties were normally distributed; and that missingness was MAR, Missing at Random. Differing and maybe unexpected results are possible, if the simulated item sets poorly measure ability levels of examinees in the population. This could be achieved by simulating items allotted to examinees with skewed distribution of abilities or simulating situations where there is a lot of missing data as a result of non-response, instead of missing by design. It will also be interesting to perform a similar study which does not only look at mean differences between subpopulations but takes into consideration correlations between covariates and achievement. Another approach could be to examine whether shorter tests can be compensated for by larger samples of test takers and vice versa; or simply, using different IRT models (e.g. the 2-PL) or missing data techniques (e.g. FIML).

7.5 General conclusion

This project examined person and item parameter recovery in different booklet designs, taking into consideration test length, sample size, item-person match, and policy-relevant subgroups within the population. Generally, for a sample size of at least 3000 students and 100 items, the results show accurate recovery of person and item parameters in all booklet designs and conditions investigated. This is true even considering parameter recovery at subgroup or subpopulation levels. Test length, sample size, and item-person match are found to be related to parameter recovery efficiency, with their effect diminishing with increasing sample size. These results are important to test developers and measurement experts, as they show that during the construction of large-scale assessments (where the sample size is typically usually over 3,000 test takers) there is less need for concern when the distribution of person abilities fail to match the distribution of item difficulties, since this does not significantly affect the precision with which person and item parameters are recovered. On the other hand, the results are beneficial to policy makers and other stakeholders in Education, since first, they prove that when using the booklet designs investigated (with samples of at least 3,000 and a test length of at least 100 items), population parameters like the mean and variance of the distribution of person abilities, are recovered accurately—both at the global population level, and for policy relevant subgroups within populations. Given growing policy concerns to ensure equity between subgroups within educational systems, the results support using matrix booklet designs as a suitable technique for estimating performance gaps between such groups. In addition, the results back using multi-matrix booklet designs as a reliable test abridgement technique in large scale assessments—which can result in great savings of material and financial resources, and lesser response burden on test takers. That notwithstanding, item-position effects were not completely considered while carrying out the studies in this project; and, different or unexpected results could be obtained if similar studies are performed with conditions involving items that poorly measure student abilities (e.g., with students having skewed ability distributions); or, simulating conditions where there is a lot of missing data due to non-response, instead of just missing by design. A similar study can be carried out which examines correlations between covariates and

achievement; or the extent to which shorter tests can compensate for large samples and vice versa.

References

Ackerman, T. A. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests Are Measuring. *Applied Measurement in Education*, 7(4), 255–278. https://doi.org/10.1207/s15324818ame0704_1

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, 21(1), 1–23. https://doi.org/10.1177/0146621697211001

Akour, M., & Al Omari, H. (2013). Empirical Investigation of The Stability of IRT Item-Parameters Estimation. *International Online Journal of Educational Sciences*, *5*(2), 291–301.

Albano, A. D. (2013). Multilevel Modelling of Item Position Effects. *Journal of Educational Measurement*, 50(4), 408–426. https://doi.org/10.1111/jedm.12026

Ali, A. M. G., Dawson, S.-J., Blows, F. M., Provenzano, E., Ellis, I. O., Baglietto, L., … Pharoah, P. D. (2011). Comparison of Methods for Handling Missing Data on Immunohistochemical Markers in Survival Analysis of Breast Cancer. *British Journal of Cancer*, 104(4), 693–699. https://doi.org/10.1038/sj.bjc.6606078

Allison, P. D. (2000). Multiple Imputation for Missing Data. *Sociological Methods & Research*, 28(3), 301–309. https://doi.org/10.1177/0049124100028003003

Allison, P. D. (2002). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Allison, P. D. (2009). Missing Data. In *The Sage Handbook of Quantitative Methods in Psychology* (pp. 72–90). Sage Publications Ltd. https://doi.org/10.4135/9780857020994.n4

Allison, P.D. (2001). Missing Data. Number 07-136 in *Sage University Papers Series on Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.

Andersen, E. B. (1972). The Solution of a Set of Conditional Estimation Equations. *Journal of the Royal Statistical Society*, 34, 42-54.

Anderson, J. O., Chiu, M. H. & Yore, L. D. (2010). First Cycle of PISA (2000–2006) – International Perspectives on Successes and Challenges: Research and Policy Directions. *International Journal of Science and Mathematics Education*, 8(3), 373–388. https://doi.org/10.1007/s10763-010-9210-y

Andrich, D. (1978). A Rating Formulation for Ordered Response Categories. *Psychometrika*, *43*, 561–573.

Andrich, D. (1979). A Model for Contingency Tables Having an Ordered Response Classification. *Biometrics*, 35(2), 403. https://doi.org/10.2307/2530343

APA Task Force on Socioeconomic Status. (2007). *Report of the APA Task Force on Socioeconomic Status*. Washington, DC: APA.

Asseburg, R., & Frey, A. (2013). Too Hard, Too Easy, or Just Right? The Relationship Between Effort or Boredom and Ability-Difficulty Fit. *Psychological Test and Assessment Modeling*, *55*(1), 92–104.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker.

Baker, F. B. (2001). *The Basics of Item Response Theory*. Madison Wisconsin: ERIC Clearing House on Assessment and Evaluation.

Baker, F. B., & Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques*. Boca Raton: CRC Press.

Baker, F. B., & Kim, S. (2017). *The Basics of Item Response Theory Using R*. Cham: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-54205-8

Baraldi, A. N., & Enders, C. K. (2010). An Introduction to Modern Missing Data Analyses. *Journal of School Psychology*, 48(1), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Bartlett, M. S., & Kendall, D. G. (1946). The Statistical Analysis of Variance Heterogeneity and the Logarithmic Transformation. *Journal of the Royal Society*, (Supplement 8), 128-138.

Beaton A.E., Barone J.L. (2017) Large-Scale Group-Score Assessment. In: Bennett R., von Davier M. (eds) *Advancing Human Assessment. Methodology of Educational Measurement and Assessment* (pp. 233-284). Springer, Cham. https://doi.org/10.1007/978-3-319-58689-2_8

Beaton, A. E. (1987). *Implementing the New Design: The NAEP 1983/1984 Technical Report*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*(2), 95-109. https://doi.org/10.2307/1165164

Béguin, A. A., & Glas, C. A. (2001). MCMC Estimation and Some Model-Fit Analysis of Multidimensional IRT Models. *Psychometrika*, *66*(4), 541-561. https://doi.org/10.1007/BF02296195

Bejar, I. I. (2014). Past and Future of Multistage Testing in Educational Reform. In: D. Yan, A. A. von Davier, and C. Lewis (Eds), *Computerized Multistage Testing: Theory and Applications* (pp. 423–438). Boca Raton, FL: CRC Press.

Belisle, M., Cassity, E., Kacilala, R., Seniloli, M. T., & Taoi, T. (2016). *Pacific Islands Literacy and Numeracy Assessment: Collaboration and Innovation in Reporting and Dissemination. Using Assessment Data in Education Policy and Practice: Examples from The Asia Pacific*. Melbourne: ACER and Bangkok: UNESCO.

Bell, R., & Lumsden, J. (1980). Test Length and Validity. *Applied Psychological Measurement*, *4*(2), 165–170. https://doi.org/10.1177/014662168000400203

Benavot, A. & Tanner, E. (2007). *The Growth of National Learning Assessments in The World, 1995–2006*. Paper Commissioned for The EFA Global Monitoring Report 2008, Education for All By 2015: Will We Make It? (pp. 1–17). UNESCO, Paris.

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Improvement of Measurement Efficiency in Multistage Tests by Targeted Assignment. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00001

Birnbaum, A. (1968*).* Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. *In F. M. Lord & M. R. Novick (Eds.), Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: MIT Press.

Bock, R. D. (1972). Estimating Item Parameters and Latent Ability When Responses are Scored in Two or More Nominal Categories. *Psychometrika*, 37(1), 29–51. https://doi.org/10.1007/bf02291411

Bock, R. D. (1997). A Brief History of Item Theory Response. *Educational Measurement: Issues and Practice*, 16(4), 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Bock, R. D., & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: An Application of The EM Algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting A Response Model for Dichotomously Scored Items. *Psychometrika*, *35*, 179–187.

Bock, R. D., & Moustaki, I. (2007). 15 Item Response Theory in a General Framework. In *Handbook of Statistics* (pp. 469–513). Elsevier. https://doi.org/10.1016/s0169-7161(06)26015-2

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic Status and Child Development. *Annual Review of Psychology*, 53, 371–399. https://doi.org/10.1146/annurev.psych.53.100901.135233

Braun, H., & von Davier, M. (2017). The Use of Test Scores from Large-Scale Assessment Surveys: Psychometric and Statistical Considerations. *Large-Scale Assessments in Education*, 5(1). https://doi.org/10.1186/s40536-017-0050-x

Braun, H., Kanjee, A. & Bettinger, E. (2006). *Improving Education, Through Assessment, Innovation, and Evaluation*. American Academy of Arts and Sciences, Cambridge, MA.

Bronner, E. (November 8, 1997). Colleges Look for Answers to Racial Gaps in Testing. *New York Times,* pp. A1, A12.

Burisch, M. (1997). Test Length and Validity Revisited. *European Journal of Personality,* 11, 303-315. https://doi.org/10.1002/(SICI)1099-0984(199711)11:4%3C303::AID-PER292%3E3.0.CO;2-%23

Camara, W. J., & Schmidt, A. E. (1999). *Group Differences in Standardized Testing and Social Stratification*. Report No. 99-5. College Entrance Examination Board.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for Quantitative Assessment of Items in Developing Patient-Reported Outcome Measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Carlson, J. E., & von Davier, M. (2013). Item Response Theory. In Bennett, R. & von Davier, M. (Eds) *Advancing Human Assessment: The Methodological, Psychological and Policy Contributions of ETS,* pp. 133-178. Cham: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-58689-2

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for The R Environment. *Journal of Statistical Software*, *48*(6), 1-29.

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, 84(4), 487–508. https://doi.org/10.3102/0034654314532697

Chen, Y., Li, X., & Zhang, S. (2018). Joint Maximum Likelihood Estimation for High-Dimensional Exploratory Item Factor Analysis. *Psychometrika*, 84(1), 124–146. https://doi.org/10.1007/s11336-018-9646-5

Chipperfield, J. O., Barr, M. L., & Steel, D. G. (2018). Split Questionnaire Designs: collecting only the data that you need through MCAR and MAR designs. *Journal of Applied Statistics*, *45*(8), 1465-1475. https://doi.org/10.1080/02664763.2017.1375085

Chuah, S. C., Drasgow F., & Luecht, R. (2006). How Big is Big Enough? Sample Size Requirements for Cast Item Parameter Estimation. *Applied Measurement in Education, 19*(3), 241–255. https://doi.org/10.1207/s15324818ame1903_5

Cochran, W. G., & Cox, G. M. (1992). *Experimental Designs* (2nd ed.). Oxford, England: Wiley.

Cohen, J. D. (1988). *Statistical Power Analysis for the Behavioural Sciences.* New York: Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. D., & Jiang, T. (1999). Comparison of Partially Measured Latent Traits Across Normal Populations. *Journal of the American Statistical Association, 94*(448), 1035–1044.

Cordero, J. M., Christobal, V., & Santin, D. (2017). Casual inference on Educational Policies: A Survey of Empirical Studies Using PISA, TIMSS and PIRLS. *MPRA Paper No. 76295*. Retrieved from https://mpra.ub.uni-muenchen.de/76295/3/MPRA_paper_76295.pdf

Coste, J., Guillemin, F., Pouchot, J., & Fermanian, J. (1997). Methodological Approaches to Shortening Composite Measurement Scales. *Journal of Clinical Epidemiology*, *50*(3), 247–252. https://doi.org/10.1016/S0895-4356(96)00363-0

Cousineau, D. & Allan, T. A. (2015). Likelihood and its use in Parameter Estimation and Model Comparison. *Mesure et évaluation en éducation*, 37 (3), 63–98. https://doi.org/10.7202/1036328ar

Cresswell, J., U. Schwantner and C. Waters (2015). *A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data*. PISA, The World Bank, Washington, D.C./OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264248373-en

Crichton, S. L. (2016). *Methods for Handling Missing Data in A Population-Based Cohort Study*. Ph.D. thesis King's College London. https://kclpure.kcl.ac.uk/portal/files/79631574/2016_Crichton_Siobhan_0948759_ethesis.pdf

Crotts, K., Zenisky, A., Sireci, S., & Li, X. (2013). Estimating Measurement Precision in Reduced-Length Multi-Stage Adaptive Testing. *Journal of Computerized Adaptive Testing*, 1(4), 67–87. https://doi.org/10.7333/1309-0104067

Custer, M. (2015). *Sample Size and Item Parameter Estimation Precision When Utilizing the One-Parameter "Rasch" Model*. Paper Presented at the Annual Meeting of the Mid-Western Educational Research Association Evanston, Illinois October 21-24, 2015.

De Ayala, & Hertzog (1991). The Assessment of Dimensionality for Use in Item Response Theory. *Multivariate Behavioral Research*, 26, 765-792. https://doi.org/10.1207/s15327906mbr2604_9

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item Parameter Recovery for the Nominal Response Model. *Applied Psychological Measurement*, 23(1), 3–19. https://doi.org/10.1177/01466219922031130

Debeer, D., & Janssen, R. (2013). Modeling Item-Position Effects Within an IRT Framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

DeMars, C. E. (2003). Sample Size and The Recovery of Nominal Response Model Item Parameters. *Applied Psychological Measurement*, *27*(4), 275-288. https://doi.org/10.1177%2F0146621603027004003

DeMars, C. E. (2016). Partially Compensatory Multidimensional Item Response Theory Models: Two Alternate Model Forms. *Educational and Psychological Measurement*, *76*(2), 231-257. https://doi.org/10.1177%2F0013164415589595

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological),* 39(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dolin, J. (2011). *Using large-scale test results for pedagogical results*. In C. Bruguière, A. Tiberghien & P. Clément (Eds.), E-Book Proceedings of the ESERA 2011 Conference: Science Learning and Citizenship. Part 10: Evaluation and Assessment of Student Learning. Lyon, France: European Science Education Research Association. Retrieved from https://www.esera.org/conference-proceedings/21-esera-2011/297-strand-10

Dong, Y., & Peng, C.-Y. J. (2013). Principled Missing Data Methods for Researchers. *SpringerPlus*, *2*(1). https://doi.org/10.1186/2193-1801-2-222

Draney, K., Wilson, M., Glück, J., & Spiel, C. (2008). Mixture Models in A Developmental Context. In G. R. Hancock, & K. M. Samuelson (Eds.), *Latent Variable Mixture Models* (pp. 199-216). Charlotte, NC: Information Age Publishing

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement*. Frankfurt am Main: Peter lang GmbH. https://doi.org/10.3726/978-3-653-04844-5

Elliot, C. D.(1990). *Differential Ability Scales (DAS)*. New York: Psychological Corporation.

Elliot, C. D., Smith, P., & McCulloch, K.(1996). *British Ability Scales II (BAS II)*. Windsor, England: NFER-Nelson.

Embretson, S. E. (1996). *The New Rules of Measurement. Psychological Assessment, 8(4), 341–349.* https://doi.org/10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Enders, C. K. (2004). The Impact of Missing Data on Sample Reliability Estimates: Implications for Reliability Reporting Practices. *Educational and Psychological Measurement*, *64*(3), 419-436. https://doi.org/10.1177/0013164403261050

Enders, C. K. (2008). A Note on the Use of Missing Auxiliary Variables in Full Information Maximum Likelihood-Based Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(3), 434–448. https://doi.org/10.1080/10705510802154307

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY, US: Guilford Press

Feinberg, R. A., & Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Finch, H., & Edwards, J. M. (2015). Rasch Model Parameter Estimation in the Presence of a Nonnormal Latent Trait Using a Nonparametric Bayesian Approach. *Educational and Psychological Measurement*, 76(4), 662–684. https://doi.org/10.1177/0013164415608418

Fischer, G. H. (1995). The Linear Logistic Test Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 131–155). New York: Springer Verlag.

Flay, B. R., Miller, T. Q., Hedeker, D., Siddiqui, O., Britton, C. F., Brannon, B. R., et al. (1995). The Television, School, and Family Smoking Prevention and Cessation Project: VIII. Student outcomes and mediating variables. *Preventive Medicine*, 24, 29 – 40.

Foley, B. P. (2010). Improving IRT Parameter Estimates with Small Sample Sizes: Evaluating the Efficacy of a New Data Augmentation Technique. Published PhD dissertation, University of Nebraska. *Public Access Theses and Dissertations from the College of Education and Human Sciences.* Paper 75. http://digitalcommons.unl.edu/cehsdiss/75.

Frey, A. & Bernhardt, R. (2012). On the Importance of Using Balanced Booklet Designs In PISA. *Psychological Test and Assessment Modeling*, *54*(4), 397–417.

Frey, A., Hartig, J., & Rupp, A. (2009). An NCME Instructional Module on Booklet Designs in Large Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice*, *28*(3), pp. 39−53. http://doi.org/10.1111/j.1745-3992.2009.00154.x

Goldman, S. H., & Raju, N. S. (1986). Recovery of One- And Two-Parameter Logistic Item Parameters: An Empirical Study. *Educational and Psychological Measurement*, *46*(1), 11–21. https://doi.org/10.1177/0013164486461002

Goldstein, H. (2004). International Comparisons of Student Attainment: Some Issues Arising from The PISA Study. *Assessment in Education*. https://doi.org/10.1080/0969594042000304618.

Gonzalez, E., & Rutkowski, L. (2010). Principles of Multiple Matrix Booklet Designs and Parameter Recovery in Large- Scale Assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 3, 125-156.

Graham, J. W., & Schafer, J. L. (1999). On the Performance of Multiple Imputation for Multivariate Data with Small Sample Size. *Statistical Strategies for Small Sample Research, 50*, 1-27

Graham, J. W., Flay, B. R., Johnson, C. A., Hansen, W. B., Grossman, L. M., & Sobel, J. L. (1984). Reliability of Self-Report Measures of Drug Use in Prevention Research: Evaluation of The Project SMART Questionnaire Via The Test-Retest Reliability Matrix. *Journal of Drug Education*, 14, 175–193.

Graham, J. W., Johnson, C. A., Hansen, W. B., Flay, B. R., & Gee, M. (1990). Drug Use Prevention Programs, Gender, And Ethnicity: Evaluation of Three Seventh-Grade Project SMART Cohorts. *Preventive Medicine*, 19, 305–313.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, 8(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned Missing Data Designs in Analysis of Change. In L. Collins & A. Sayer (Eds.), *New Methods for The Analysis of Change* (pp. 335–353). Washington, DC: American Psychological Association.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned Missing Data Designs in Psychological Research. *Psychological Methods*, 11(4), 323–343. https://doi.org/10.1037/1082-989x.11.4.323

Greaney, V., & Kellaghan, T. (Eds.). (2008). *Assessing National Achievement Levels in Education*. Washington, DC: World Bank.

Grek, S. (2009). Governing by Numbers: The PISA 'Effect' In Europe. *Journal of Education Policy*, 24(1), 23–37.

Gressard, R., & Loyd, B. (1991). A Comparison of Item Sampling Plans In The Application Of Multiple Matrix Sampling. *Journal of Educational Measurement*, *28*(**2**), 119−130. https://doi.org/10.1111/j.1745-3984.1991.tb00348.x

Guyer, R., & Thompson, N. A. (2011). *User's Manual for Xcalibre 4.1*. St. Paul, MN: Assessment Systems Corporation.

Hahne, J. (2008). Analyzing Position Effects Within Reasoning Items Using the LLTM For Structurally Incomplete Data. *Psychology Science Quarterly*, 50, 379-390.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K. (1989). Principles and Selected Applications of Item Response Theory. In RL. Linn, (Ed.), *Educational Measurement* (3rd Ed., Pp. 143-200). New York, NY: Macmillan.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). An NCME Instructional Module on Educational Measurement: Issues and Practice, *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development*, 12(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Jones, R. W. (1994). Item Parameter Estimation Errors and Their Influence on Test Information Functions. *Applied Measurement in Education*, 7(3), 171–186. https://doi.org/10.1207/s15324818ame0703_1

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hannover, B., & Kessels, U. (2011). Sind Jungen Die Neuen Bildungsverlierer? Empirische Evidenz Für Geschlechterdisparitäten Zuungunsten Von Jungen Und Erklärungsansätze1. *Zeitschrift Für Pädagogische Psychologie*, 25(2), 89–103. https://doi.org/10.1024/1010-0652/a000039

Hansen, W. B., Johnson, C. A., Flay, B. R., Graham, J. W., & Sobel, J. L. (1988). Affective and Social Influences Approaches to The Prevention of Multiple Substance Abuse Among Seventh Grade Students: Results from Project SMART. *Preventive Medicine*, 17, 135–154.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary Variables in Multiple Imputation in Regression with Missing X: A Warning Against Including Too Many in Small Sample Research. *BMC Medical Research Methodology*, *12*, 184. https://doi.org/10.1186/1471-2288-12-184

Hartig, J., & Höhler, J. (2009). Multidimensional IRT Models for The Assessment of Competencies. *Studies in Educational Evaluation*, *35*(2-3), 57-63. https://doi.org/10.1016/j.stueduc.2009.10.002

Harwell, M. R., & Janosky, J. E. (1991). An Empirical Study of The Effects of Small Datasets and Varying Prior Variances on Item Parameter Estimation In BILOG. *Applied*

Psychological Measurement, 15(3), 279–291. https://doi.org/10.1177/014662169101500308

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment*, *7*(3), 238. https://doi.org/10.1037/1040-3590.7.3.238

He, Q., & Wheadon, C. (2013). The Effect of Sample Size on Item Parameter Estimation for The Partial Credit Model. *International Journal of Quantitative Research in Education*, *1*(3), 297-315. https://doi.org/10.1504/ijqre.2013.057692

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015a): Effects of Design Properties on Parameter Estimation in Large-Scale Assessments. In *Educational and Psychological Measurement* 75 (6), 1021-1044. https://doi.org/10.1177/0013164415573311

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015b). Modeling Booklet Effects for Nonequivalent Group Designs in Large-Scale Assessment. *Educational and Psychological Measurement*, *75*(4), 568–584. https://doi.org/10.1177/0013164414554219

Heyneman, SP, Lee, B (2014) The Impact of International Studies of Academic Achievement on Policy and Research. In: Rutkowski, L, Von Davier, M, Rutkowski, D (Eds) *Handbook of International Large-Scale Assessment*. CRC Press, Pp. 37–75

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining Item-Position Effects in Large-Scale Assessment Using the Linear Logistic Test Model. *Psychology Science Quarterly*, 50(3), 391-402.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analyzing Item Position Effects Due to Test Booklet Design Within Large-Scale Assessment. *Educational Research and Evaluation*, *17*(6), 497–509. https://doi.org/10.1080/13803611.2011.632668

Howell, D. C. (2007). The Treatment of Missing Data. In *The SAGE Handbook Of Social Science Methodology* (Pp. 212–226). SAGE Publications Ltd. https://doi.org/10.4135/9781848607958.n11

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of Two and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study. *Applied Psychological Measurement*, *6*(3), 249–260. https://doi.org/10.1177/014662168200600301

Jacob, B., & Rothstein, J. (2016). The Measurement of Student Ability in Modern Assessment Systems. *The Journal of Economic Perspectives, 30*(3), 85–107

Jencks, C., & Phillips, M. (1998). *The Black-White Test Score Gap.* Washington, DC: The Brookings Institute.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample Size Requirements for Estimation of Item Parameters in The Multidimensional Graded Response Model. *Frontiers in Psychology*, *7*, 109. https://doi.org/10.3389/fpsyg.2016.00109

Johnson, E. G. (1992). The Design of The National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 95–110. https://doi.org/10.1111/j.1745-3984.1992.tb00369.x

Johnson, M. S. (2007). contro. *Journal of Statistical Software*, *20*(10), 1-24.

Johnson, M. S. (2018). Test Information Function. In B. Frey (Ed.), *The SAGE Encyclopaedia of Educational Research, Measurement, And Evaluation*. Thousand Oaks: SAGE Publications. https://dx.doi.org/10.4135/9781506326139

Kamens, D. H., & Mcneely, C. L. (2010). Globalization and The Growth of International Educational Testing and National Assessment. *Comparative Education Review*, *54*(1), 5–25. https://doi.org/10.1086/648471

Kang, T. (2006). *Model Selection Methods for Unidimensional and Multidimensional IRT Models* (Doctoral Dissertation, University of Wisconsin-Madison). Retrieved From https://www.researchgate.net/publication/268396333

Kaplan, D., & Su, D. (2016). On Matrix Sampling and Imputation of Context Questionnaires with Implications for The Generation of Plausible Values in Large-Scale Assessments. *Journal of Educational and Behavioral Statistics*, 41(1), 57–80. https://doi.org/10.3102/1076998615622221

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children (K- Abc)*. Circle Pines, Minn.: American Guidance Service.

Kieftenbeld, V., & Natesan P. (2012). Recovery of Graded Response Model Parameters: A Comparison of Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement*, 36(5): 399-419. https://doi.org/10.1177/0146621612446170

Kirsch, I., Von Davier, M., Gonzalez, E., & Yamamoto, K. (Eds.). (2013). *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, And Educational Research*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4629-9

Klinger, D. A., & Rogers, W. T. (2011). Teachers' Perceptions of Large-Scale Assessment Programs Within Low-Stakes Accountability Frameworks. *International Journal of Testing, 11*, 122-143. https://doi.org/10.1080/15305058.2011.552748

Klinger, D.A., Deluca, C., & Miller, T. (2008). The Evolving Culture of Large-Scale Assessments in Canadian Education. *Canadian Journal of Educational Administration and Policy*, 76.

Köhler, C. (2017). *Isn't Something Missing? Latent Variable Models Accounting for Item Nonresponse* (PhD Thesis). Freie Universität Berlin. http://www.diss.fu-berlin.de/diss/servlets/mcrfilenodeservlet/fudiss_derivate_000000020773/dissertation_koehler.pdf

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, And Linking: Methods and Practices*. Springer Science & Business Media.

Kreiner, S. (2013). The Rasch Model for Dichotomous Items. In K. Christensen, S. Kreiner, M. Mesbah (Eds) *Rasch Models In Health* (Pp. 27-42). London: John Wiley & Sons, Inc.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test Length and Decision Quality in Personnel Selection: When Is Short Too Short? *International Journal of Testing*, *12*(4), 321–344. https://doi.org/10.1080/15305058.2011.643517

Kubinger, K. D., & Draxler, C. (2007). A Comparison of The Rasch Model and Constrained Item Response Theory Models for Pertinent Psychological Test Data. In Von Davier, M., & Carstensen, C. H. (Eds). *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (Pp. 293-311). New York: Springer. https://doi.org/10.1007/978-0-387-49839-3

Kubinger, K. D., & Wurst, E. (2000). *Adaptives Intelligenz Diagnosticum—Version 2.1 (AID 2).[Adaptive Intelligence Diagnosticum—Version 2.1.].* Göttingen, Germany: Beltz.

Kulpoo, D., & P. Coustère. (1999). Developing National Capacities for Assessment and Monitoring Through Effective Partnerships. In *Partnerships for Capacity Building and Quality Improvements in Education: Papers from The ADEA 1997 Biennial Meeting*, Dakar, Senegal. Paris: ADEA (Association for The Development of Education in Africa).

Laukaityte, I., & Wiberg, M. (2017). Using Plausible Values in Secondary Analysis in Large-Scale Assessments. *Communications in Statistics - Theory and Methods*, 46(22), 11341–11357. https://doi.org/10.1080/03610926.2016.1267764

Le, L. T., & Adams, R. J. (2013). *Accuracy of Rasch Model Item Parameter Estimation.* ACER Project Report. http://research.acer.edu.au/ar_misc/13

Levine, T. R., & Hullett, C. R. (2002). Eta Squared, Partial Eta Squared, And Misreporting of Effect Size in Communication Research. *Human Communication Research*, *28*(4), 612–625. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x

Lietz, P., Cresswell, J. C., Rust, K. F., & Adams, R. J. (2017). Implementation of Large-Scale Education Assessments. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.),

Implementation of Large-Scale Education Assessments (Pp. 1–25). https://doi.org/10.1002/9781118762462.ch1

Lim, R. G., & Drasgow, F. (1990). Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning. *Journal of Applied Psychology*, *75*(2), 164–174. https://doi.org/10.1037/0021-9010.75.2.164

Linacre, J. M. (1989). *Many-Faceted Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (2006). *WINSTEPS Rasch Measurement Computer Program*. Chicago: WINSTEPS. Com.

Linn, R. L. (2003). Accountability: Responsibility and Reasonable Expectations. *Educational Researcher*, 32(7), 3–13. https://doi.org/10.3102/0013189x032007003

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2 Ed.). Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley

Little, R. J., & Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research*, *18*(2-3), 292-326. https://doi.org/10.1177%2f0049124189018002004

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ.

Lord, F.M. (1962). Estimating Norms by Item Sampling. *Educational and Psychological Measurement*, 22(2), 259-267. http://dx.doi.org/10.1177/001316446202200202

Luecht, R. M., & Miller, T. R. (1992). Unidimensional Calibrations and Interpretations of Composite Traits for Multidimensional Tests, *Applied Psychological Measurement*, 16, 279-293. https://doi.org/10.1177/014662169201600308

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*, 149–174.

Masters, G. N. (1985). A Comparison of Latent-Trait and Latent-Class Analysis of Likert-Type Data. *Psychometrika*, *50*, 69–82.

Masters, G. N. (2010) *Teaching and Learning School Improvement Framework.* Melbourne : Australian Council for Educational Research (ACER); Brisbane : Department of Education and Training

Masters, G. N., & Wright, B. D. (1997). The Partial Credit Model. In W. J. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (Pp. 101–21). New York: Springer-Verlag.

Mcardle, J. J. (1994). Structural Factor Analysis Experiments with Incomplete Data. *Multivariate Behavioral Research*, *29*(4), 409–454. https://doi.org/10.1207/s15327906mbr2904_5

Mislevy, R. J. (1984). Estimating Latent Distributions. *Psychometrika*, 49(3), 359–381. https://doi.org/10.1007/bf02306026

Mislevy, R. J. (1985). Estimation of Latent Group Effects. *Journal of The American Statistical Association, 80,* 993–997.

Mislevy, R. J. (1986). Bayes Modal Estimation in Item Response Models. *Psychometrika*, 51, 177-195.

Mislevy, R. J. (1991). Randomization-Based Inference About Latent Variables from Complex Samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal Estimation Procedures. In A. E. Beaton, *Implementing the New Design: The NAEP 1983-84 Technical Report* (Pp. 293-360). (No. 15-TR-20) Princeton, NJ Educational Testing Service.

Mislevy, R. J., And Wu, P. (1996). Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, And Adaptive Testing *(ETS Research Reports Series No. RR-96-30-ONR)*. Princeton, NJ: Educational Testing Service

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating Population Characteristics from Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement, 29,* 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R.J. (1991). Randomization-Based Inference About Latent Variables from Complex Samples. *Psychometrika*, 56, 177–196.

Monseur, C., & Adams, R. (2009). Plausible Values: How to Deal with Their Limitations. *Journal of Applied Measurement*, 10 (3), 320 – 334

Montgomery, M., & Skorupski, D. W. (2012). Investigation of IRT Parameter Recovery And Classification Accuracy In Mixed Format, 27. *Paper Presented At The Annual Meeting Of The National Council Of Measurement In Education*, April 2012. Retrieved From https://aai.ku.edu/sites/cete.ku.edu/files/docs/presentations/2012_04_montgomery%20irt%20classification_ncme.pdf

Moy, M. L. Y., & Barcikowski, R. S. (1974). Item Sampling: Optimal Number of People and Items. *The Journal of Experimental Education*, *42*(3), 46–52.

Naemi, B., Gonzalez, E., Bertling, J., Betancourt, A., Burrus, J., Kyllonen, P., Minsky, J., Lietz, P., Klieme, E., Vieluf, S., Lee, J., & Roberts, R.D. (2013). Large-Scale Group Score Assessments: Past, Present, And Future. In D. Saklofske & V. Schwean (Eds.), *Oxford Handbook of Psychological Assessment of Children and Adolescents*. Cambridge, MA: Oxford University Press, Pp. 129-149

National Research Council. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials*. Committee on National

Statistics, Division of Behavioral And Social Sciences and Education. Washington, DC: The National Academies Press.

OECD (2012). Sample Design. *PISA 2009 Technical Report*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

OECD (2014), PISA 2012 *Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised Edition, February 2014), PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

Padilla, J. L., Azevedo, C. L. N., & Lachos, V. H. (2018). Multidimensional Multiple Group IRT Models with Skew Normal Latent Trait Distributions. *Journal of Multivariate Analysis*, 167, 250–268. https://doi.org/10.1016/j.jmva.2018.05.003

Palmer, H. D. (2011). Conditional Maximum Likelihood Estimation. In: Lewis-Beck M., Bryman A., & Tim L. (Eds) *The SAGE Encyclopaedia of Social Science Research Methods* (Vols. 1-0). Thousand Oaks, CA: Sage Publications, Inc. https://dx.doi.org/10.4135/9781412950589.n153

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.) (2013). *IQB-Ländervergleich 2012*. Münster: Waxmann.

Partchev, I. (2017). *Irtoys: A Collection of Functions Related to Item Response Theory*. [R Software Package, Version 0.2.1]. Available From https://cran.r-project.org/package=irtoys

Patsula, L. N., & Gessaroli M. E. (1995, April). *A Comparison of Item Parameter Estimates and Iccs Produced with TESTGRAF And BILOG Under Different Test Lengths and Sample Sizes*. Paper Presented at The Annual Meeting of The National Council on Measurement In Education, San Francisco, CA.

Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, *74*(4), 525–556. https://doi.org/10.3102/00346543074004525

Pigott, T. D. (2001) A Review Of Methods For Missing Data, *Educational Research And Evaluation*, 7:4, 353-383, https://doi.org/10.1076/edre.7.4.353.8937

Pohl, S., & Aßmann, C. (2015). Missing Values in Large-Scale Assessment Studies. *Psychological Test and Assessment Modeling*, 57(4), 469-471.

Polit, D. F., & Yang, F. M. (2014). *Measurement and The Measurement of Change: A Primer for Health Professionals*. Baltimore, MD: Lippincott, Williams, And Wilkins.

Postlethwaite, T. N. & Kellaghan, T. (2008). *National Assessments of Education Achievement*. Jointly Published By IIEP, Paris, France and IAE, Brussels, Belgium. Available At: http://unesdoc.unesco.org/images/0018/001817/181753e.pdf (Accessed 15 March 2019).

Preece, D. A. (1990). Fifty Years of Youden Squares: A Review. *Bulletin of The Institute of Mathematics and Its Applications, 26,* 65−75.

Price, L. R. (2017). *Psychometric Methods: Theory into Practice*. New York: The Guilford Press.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. http://www.r-project.org

Raghunathan, T., & Grizzle, J. (1995) A Split Questionnaire Survey Design, *Journal of The American Statistical Association*, 90(429), 54-63. https://doi.org/10.1080/01621459.1995.10476488

Rasch, G. (1960). *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. Oxford, England: Nielsen & Lydiche.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press (Original Work Published 1960).

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, 4, 207-230. https://doi.org/10.3102/10769986004003207

Reckase, M. D. (1997). The Past and Future of Multidimensional Item Response Model. *Applied Psychological Measurement*, 21, 25-36. https://doi.org/10.1177/0146621697211002

Reynolds, T., Pekins, K., & Brutten, S. (1994). A Comparative Item Analysis Study of a Language Testing Instrument. *Applied Psychological Measurement*, 11, 1-13. https://doi.org/10.1177/026553229401100102

Rhemtulla, M., & Hancock, G. R. (2016). Planned Missing Data Designs in Educational Psychology Research. *Educational Psychologist*, *51*(3–4), 305–316. https://doi.org/10.1080/00461520.2016.1208094

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test Analysis Modules*. R Package Version 1.995-0, [Computer Software]. https://cran.r-project.org/package=tam

Roscigno, V. J., & Crowle, M. L. (2009). Rurality, Institutional Disadvantage, And Achievement/Attainment*. *Rural Sociology*, 66(2), 268–292. https://doi.org/10.1111/j.1549-0831.2001.tb00067.x

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* [Textbook Test Theory, Test Construction] (2nd Ed.). Psychologie Lehrbuch. Bern: Verlag Hans Huber.

Roth, P. L. (1994). Missing Data: A Conceptual Review for Applied Psychologists. *Personnel Psychology*, 47(3), 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1977). Assignment to Treatment Group on The Basis of a Covariate. *Journal of Educational Statistics*, 2(1), 1–26. https://doi.org/10.3102/10769986002001001

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rutkowski, L., Gonzalez, E., Von Davier, M., & Zhou, Y. (2014). Assessment Design for International Large-Scale Assessments. In L. Rutkowski, M. Von Davier & D. Rutkowski (Eds) *Handbook of International Large-Scale Assessment,* Pp.75-96. Boca Raton: CRC Press.

Sacks, P. (1997). Standardized Testing: Meritocracy's Crooked Yardstick. *Change,* Pp. 25–31.

Şahin, A., & Anıl, D. (2016). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, *17*(1). https://doi.org/10.12738/estp.2017.1.0270

Samejima, F. (1969). Estimation of Latent Ability Using A Response Pattern of Graded Scores. *Psychometrika Monograph Supplement*. *34*(4, Pt. 2), 100.

Schafer, J (1997). Analysis of Incomplete Multivariate Data. *Monographs on Statistics and Applied Psychology 72*. London: Chapman & Hall/CRC.

Schafer, J. L. (2003). Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Neerlandica, 57*(1), 19-35

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of The State of The Art. *Psychological Methods, 7*(2), 147-177. https://doi.org/10.1037/1082-989x.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, 33(4), 545–571. https://doi.org/10.1207/s15327906mbr3304_5

Schleicher, A., Zimmer, K., Evans, J., & Clements, N. (2009). PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science. *OECD Publishing (NJ1)*.

Seastrom, M. (2017). *Best Practices for Determining Subgroup Size in Accountability Systems While Protecting Identifiable Student Information. (IES 2017-147).* U.S. Department of Education, Institute of Education Sciences. Washington, DC. Retrieved July 12, 2018 From https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2017147

Shin, A.-Y. (2016). *Investigating the Effects of Missing Data Treatments on Item Response Theory Vertical Scaling*. PhD Thesis, University of Iowa, 2016. https://doi.org/10.17077/etd.g0o7zsk9

Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, MA: Ballinger Publishing Company.

Si, C.-F. (2002). *Ability Estimation Under Different Item Parameterization and Scoring Models*. Doctoral Dissertation, University of North Texas [Online]. Retrieved From http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.852.2227&rep=rep1&type=pdf

Si, C.-F., & Schumacker, R. E. (2004). Ability Estimation Under Different Item Parameterization and Scoring Models. *International Journal of Testing*, *4*(2), 137–181. https://doi.org/10.1207/s15327574ijt0402_3

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned Missing-Data Designs in Experience-Sampling Research: Monte Carlo Simulations of Efficient Designs for Assessing Within-Person Constructs. *Behavior Research Methods*, *46*(1), 41–54. https://doi.org/10.3758/s13428-013-0353-y

Simon, M. (Ed.), Ercikan, K. (Ed.), Rousseau, M. (Ed.). (2013*). Improving Large-Scale Assessment in Education*. New York: Routledge.

Simpson, M., Gong, B., & Marion S. (2006). *Effect of Minimum Cell Sizes and Confidence Interval Sizes for Special Education Subgroups on School-Level AYP Determinations* (Synthesis Report 61). Minneapolis, MN: University of Minnesota, National Center On Educational Outcomes. Retrieved From https://conservancy.umn.edu/bitstream/handle/11299/173884/synthesis61.pdf?sequence=1&isallowed=y

Sinharay, S., Stern, H.S., & Russell, D. (2001). The Use of Multiple Imputation for The Analysis of Missing Data. *Psychological Methods*, 6(4), 317-329. https://doi.org/10.1037/1082-989x.6.4.317

Spearman, C. (1910). Correlation Calculated from Faulty Data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steenkamp, J.-B. E. M., De Jong, M. G., & Baumgartner, H. (2010). Socially Desirable Response Tendencies in Survey Research. *Journal of Marketing Research*, 47(2), 199–214. https://doi.org/10.1509/jmkr.47.2.199

Stone, C. A. (1992). Recovery of Marginal Maximum Likelihood Estimates in The Two-Parameter Logistic Response Model: An Evaluation of Multilog. *Applied Psychological Measurement, 16*(1), 1–16. https://doi.org/10.1177/014662169201600101

Stone, M., & Yumoto, F. (2004). The Effect of Sample Size for Estimating Rasch/IRT Parameters with Dichotomous Items. *Journal of Applied Measurement*, *5*(1), 48–61.

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., ... & Kunze, K. L. (2013). Designing Small-Scale Tests: A Simulation Study of Parameter Recovery with the 1-PL. *Psychological Test and Assessment Modeling*, *55*(4), 335.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of Parameters in The Three-Parameter Latent Trait Model. In *New Horizons in Testing* (Pp. 13-30). Academic Press. https://doi.org/10.1016/b978-0-12-742780-5.50009-3

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small Sample Estimation in Dichotomous Item Response Models: Effect of Priors Based on Judgmental Information on The Accuracy of Item Parameter Estimates. *Applied Psychological Measurement*, *27*(1), 27–51. https://doi.org/10.1177/0146621602239475

Tang, K. L., Way, W. D., & Carey, P. A. (1993). *The Effect of Small Calibration Sample Sizes on TEOFL IRT-Based Equating (TOEFL Technical Report TR-7)*. Princeton, NJ: Educational Testing Service.

Tay, L., Huang, Q., & Vermunt, J. K. (2016). Item Response Theory with Covariates (IRT-C): Assessing Item Recovery and Differential Item Functioning for The Three-Parameter Logistic Model. *Educational and Psychological Measurement*, *76*(1), 22–42. https://doi.org/10.1177/0013164415579488

Taylor, B. J., Graham, J. W., Palmer, R. F., & Tatterson, J. W. (1998, June). *Interpreting Latent Variable Models Involving Self-Report and Objective Measures*. Paper Presented at The Annual Meeting of The Society for Prevention Research, Park City, UT.

Tellegen, A., & Briggs, P. F. (1967). Old Wine in New Skins: Grouping Wechsler Subtests into New Scales. *Journal of Consulting Psychology, 31*(5), 499-506. http://dx.doi.org/10.1037/h0024963

Thissen, D., & Wainer, H. (1982). Some Standard Errors in Item Response Theory. *Psyhometrika*, 47(4), 397-412. https://doi.org/10.1007/bf0227305

Thomas, N., & Gan, N. (1997). Generating Multiple Imputations for Matrix Sampling Data Analysed with Item Response Models. *Journal of Educational and Behavioural Statistics*, 22(4), 425–445. https://doi.org/10.3102/10769986022004425

Thomas, N., Raghunathan, T., Schenker, N., Katzoff, M., & Johnson, C. (2006). An Evaluation of Matrix Sampling Methods Using Data from The National Health and Nutrition Examination Survey. *Survey Methodology*, 32(2), 217-231

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). *Using Large-Scale Assessments of Students' Learning to Inform Education Policy: Insights from The*

Asia-Pacific Region. Melbourne, Bangkok: Australian Council for Educational Research, UNESCO.

Toland, M. (2008). Determining the Accuracy of Item Parameter Standard Error of Estimates In BILOG-MG 3. Published PhD Dissertation, University of Nebraska. *Open Access Theses and Dissertations from The College of Education and Human Sciences.* Paper 22. Retrieved From http://digitalcommons.unl.edu/cehsdiss/22

Tolonen, H. (2005). *Towards the High Quality of Population Health Surveys: Standardization and Quality Control*. KTL (Kansanterveyslaitos Folkhälsoinstitutet), Helsinki.

Traub, R. E. (1983). A Priori Considerations in Choosing an Item Response Model. In R. K. Hambleton (Ed.), *Applications of Item Response Theory* (Pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Van Der Linden, W. J., & Boekkooi-Timminga, E. (1989). A Maximin Model For IRT-Based Test Design with Practical Constraints. *Psychometrika*, 54(2), 237–247. https://doi.org/10.1007/bf02294518

Van Ginkel, J. R., Linting, M., Rippe, R. C. A., & Van Der Voort, A. (2019). Rebutting Existing Misconceptions About Multiple Imputation as A Method for Handling Missing Data. *Journal of Personality Assessment*, 1–12. https://doi.org/10.1080/00223891.2018.1530680

von Davier, M. (2003). Comparing Conditional and Marginal Direct Estimation of Subgroup Distributions. *ETS Research Report Series,* 2003(1), i-36. https://doi.org/10.1002/j.2333-8504.2003.tb01894.x

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What Are Plausible Values and Why Are They Useful. *IERI Monograph Series*, *2*(1), 9-36.

Von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (Eds.). (2013). *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, And Educational Research*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4629-9

Wagemaker, H. (2014). International Large-Scale Assessments: From Research to Policy. In Rutkowski, Von Davier, & Rutkowski (Eds.), Handbook *of International Large-Scale Assessments Background, Technical Issues and Methods of Data Analyses*. Boca Raton: CRC Press.

Wang, W.-C., & Chen, C.-T. (2005). Item Parameter Recovery, Standard Error Estimates, And Fit Statistics of The Winsteps Program for The Family of Rasch Models. *Educational and Psychological Measurement*, *65*(3), 376–404. https://doi.org/10.1177/0013164404268673

Weirich, S., Hecht, M., & Bohme, K. (2014). Modeling Item Position Effects Using Generalized Linear Mixed Models. *Applied Psychological Measurement*, *38*(7), 535–548. Https://Doi.Org/10.1177/0146621614534955

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Appl. Psychol. Meas.* 6, 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J., & Minden, S. V. (2012). *A Comparison of Item Parameter Estimates from Xcalibre 4.1 And Bilog-MG.* St. Paul, MN: Assessment Systems Corporation.

White, I. R., Royston, P., & Wood, A. M. (2010). Multiple Imputation Using Chained Equations: Issues and Guidance For Practice. *Statistics in Medicine,* 30(4), 377–399. https://doi.org/10.1002/sim.4067

Wilcox, R. R. (1980). Determining the Length of a Criterion-Referenced Test. *Applied Psychological Measurement*, 4(4), 425–446. https://doi.org/10.1177/014662168000400402

Wilson, M. R.(1989). Saltus: A Psychometric Model of Discontinuity in Cognitive Development. *Psychological Bulletin*, *105*, 276–289.

Wingersky, M., Kaplan, B.A., & Beaton, A.E. (1987). "Joint Estimation Procedures" In A. E. Beaton (Ed*.), Implementing the New Design: The NAEP 1983–84 Technical Report* (Pp.285– 92) (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Wiseman, A. W. (2010). The Uses of Evidence for Educational Policymaking: Global Contexts and International Trends. In: Luke, A., Kelly, G. J. & Green, J. (Eds.) *Review of Research in Education* (Vol. 34, Pp. 1–24). American Educational Research Association, Washington, DC.

Wolf, A. (2006). *Shorter Tests Through the Adaptive Use of Planned Missing Data in Sampling Designs.* Ph.D. Thesis, University of Jena. [Online]. Retrieved From www.dbthueringen.de/servlets/derivateservlet/derivate-11463

Wollack, J. A., Sung, H. J., & Kang, T. (2006, April). The Impact of Compounding Item Parameter Drift on Ability Estimation. In *Annual Meeting of The National Council on Measurement in Education, San Francisco, CA* (Vol. 147).

Wood, J., Matthews, G. J., Pellowski, J., & Harel, O. (2018). Comparing Different Planned Missingness Designs in Longitudinal Studies. *Sankhya B*. https://doi.org/10.1007/s13571-018-0170-5

Wright, B. D., & Masters, G. N.(1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright, B. D., & Mok, M. M. C. (2004). *An Overview of The Family of Rasch Measurement Models*. In E. V. Smith, & R. M. Smith (Eds) Introduction to Rasch Measurement: Theory, Models and Applications (Pp. 1-24). Maple Grove: JAM Press.

Wu, M., & Adams, R. (2007). *Applying the Rasch Model to Psycho-Social Measurement: A Practical Approach*. Melbourne: Educational Measurement Solutions.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers: Theory into Practice*. Singapore: Springer Nature Pte Ltd. https://doi.org/10.1007/978-981-10-3302-5

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Singapore: Springer Nature Pte Ltd. https://doi.org/10.1007/978-981-10-3302-5

Yamamoto, K., & Kulick, E. (2000). Scaling Methodology and Procedures for The TIMSS Mathematics and Science Scales. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 Technical Report*. Chestnut Hill, MA: Boston College.

Yamamoto, K., Khorramdel, L., & Von Davier, M. (2013, Updated 2016). Scaling PIAAC Cognitive Data. In *OECD (2013), Technical Report of The Survey of Adult Skills (PIAAC)*, Chapter 17 (Pp. 406-438), PIAAC, OECD Publishing. Retrieved From https://www.oecd.org/skills/piaac/technical_report_2nd_edition_chapters_17-23.pdf

Yang, F. M., & Kao, S. T. (2014). Item Response Theory for Measurement Validity. *Shanghai Archives of Psychiatry*, 26(3), 171-177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons from Machine Learning. *Perspect Psychol Sci* 12(6): 1100-1122. https://doi.org/10.1177/1745691617693393

Yen, W. M. (1987). A Comparison of The Efficiency and Accuracy of Bilog And Logist. *Psychometrika*, *52*(2), 275–291. https://doi.org/10.1007/bf02294241

Yoes, M. (1995). *An Updated Comparison of Micro-Computer-Based Item Parameter Estimation Procedures Used with the 3-Parameter IRT Model*. Saint Paul, MN: Assessment Systems Corporation.

Younger, M. & Warrington, M (2005). *Raising Boys' Achievement*. Research Report RR636. University of Cambridge: Faculty of Education.

Yousfi, S. (2005). Mythen Und Paradoxien Der Klassischen Testtheorie (I). *Diagnostica*, 51(1), 1–11. https://doi.org/10.1026/0012-1924.51.1.1

Yousfi, S., & Böhme, H. (2012). Principles and Procedures of Considering Context Effects in The Development of Calibrated Item Pools: Conceptual Analysis and Empirical Illustration. *Psychological Test and Assessment Modeling*, 54, 366-396.

Yuan, K.-H., & Bentler, P. M. (2000). 5. Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. *Sociological Methodology*, 30(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An Application of Item Response Theory to Psychological Test Development. *Psicologia: Reflexão E Crítica*, 29(1). https://doi.org/10.1186/s41155-016-0040-x

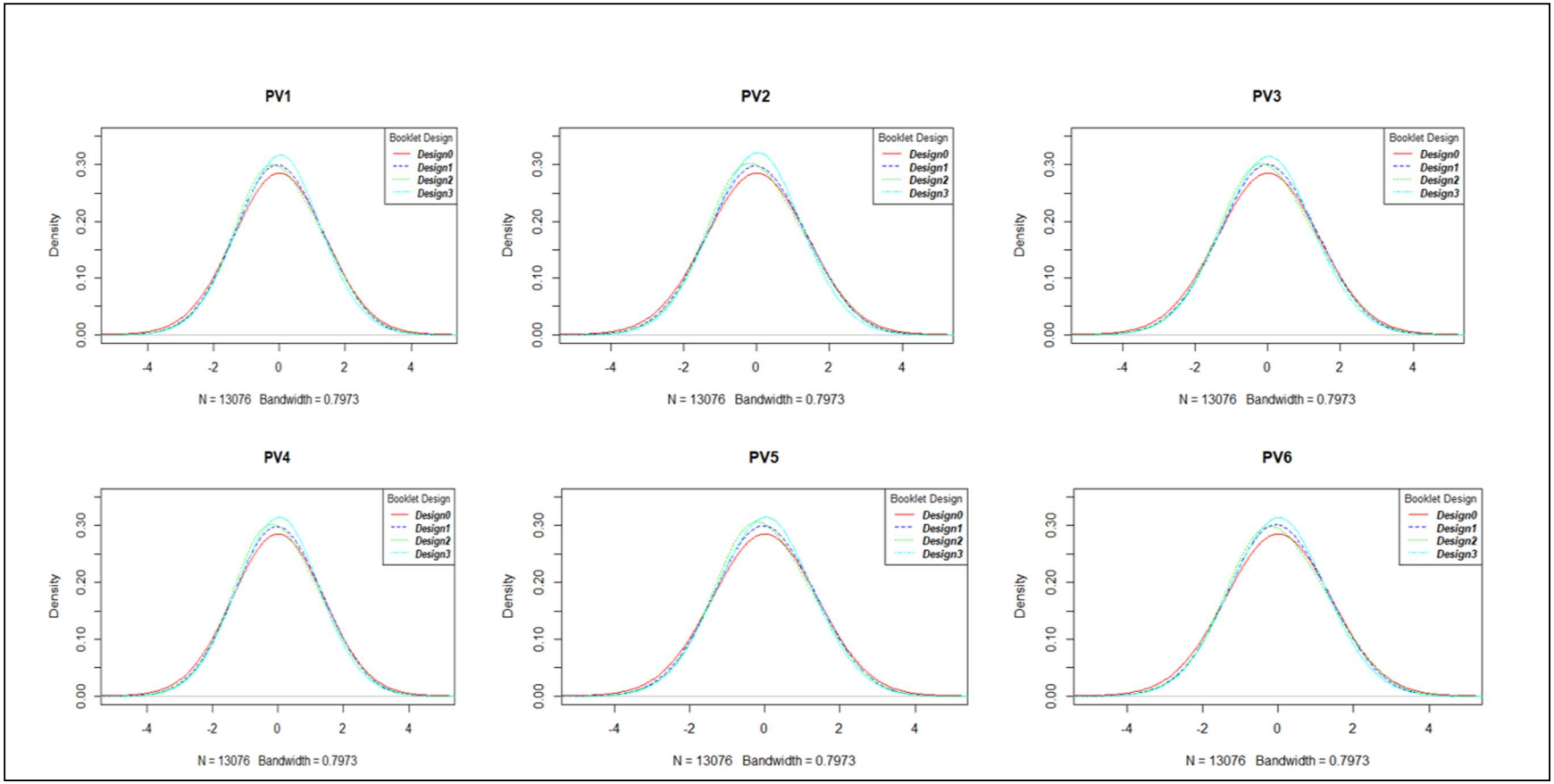
Zięba, A. (2013). The Item Information Function in One and Two-Parameter Logistic Models – A Comparison and Use in The Analysis of The Results of School Tests. *Didactics of Mathematics*, 10(14), 87-96. https://doi.org/10.15611/dm.2013.10.08

Zwinderman, A. H. And Van Der Wollenberg, A. L. (1990). Robustness of Marginal Maximum Likelihood Estimation in The Rasch Model. *Applied Psychological Measurement*, 14, 73-81.

Appendix A Additional Results

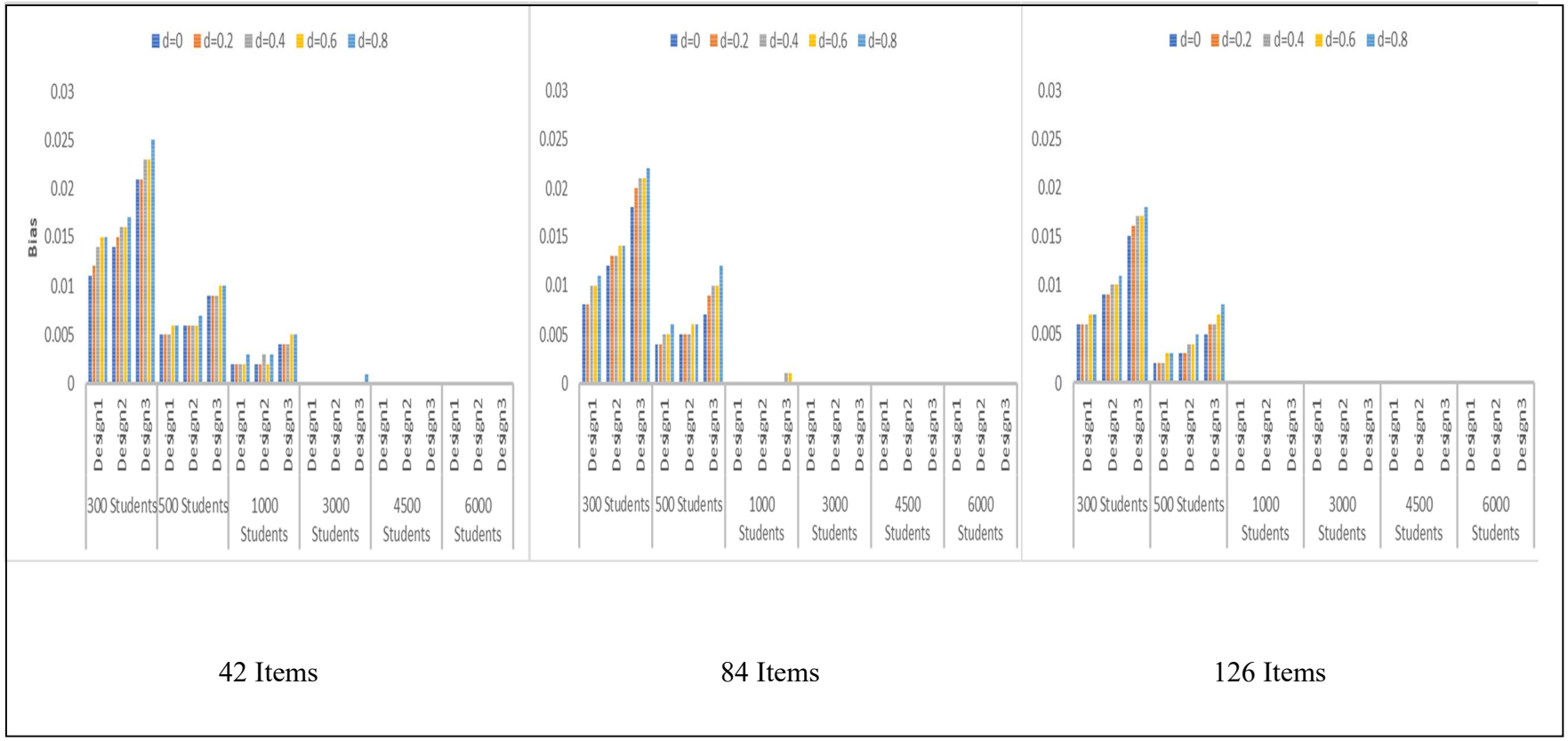
This section of the appendix gives additional results of the empirical and simulation studies performed in chapter five and chapter six of this PhD project.

A.1 Recovery of person ability distributions across various booklet designs for the set of first six plausible values



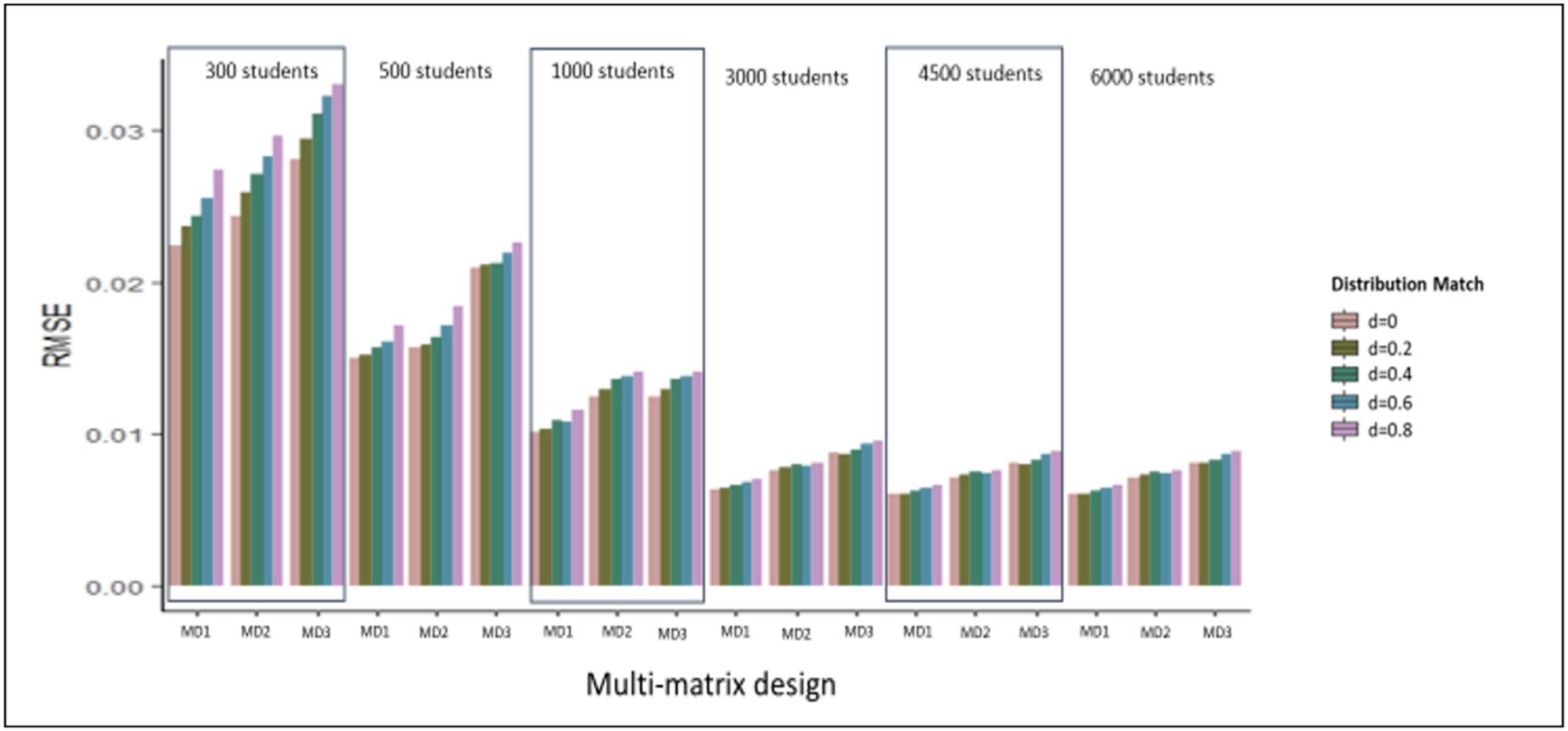
Note. The figure shows the distribution of person ability estimates recovered from the various booklet designs using the first six sets of plausible values (denoted by PV1 to PV6 respectively). Design0 contains the complete dataset with no missing data while the designs get sparser from Design1 to Design3. Six plausible values for used for brevity and especially as the results don't change much across the various PVs.

A.2 Bias of recovered mean person ability across item-person match conditions



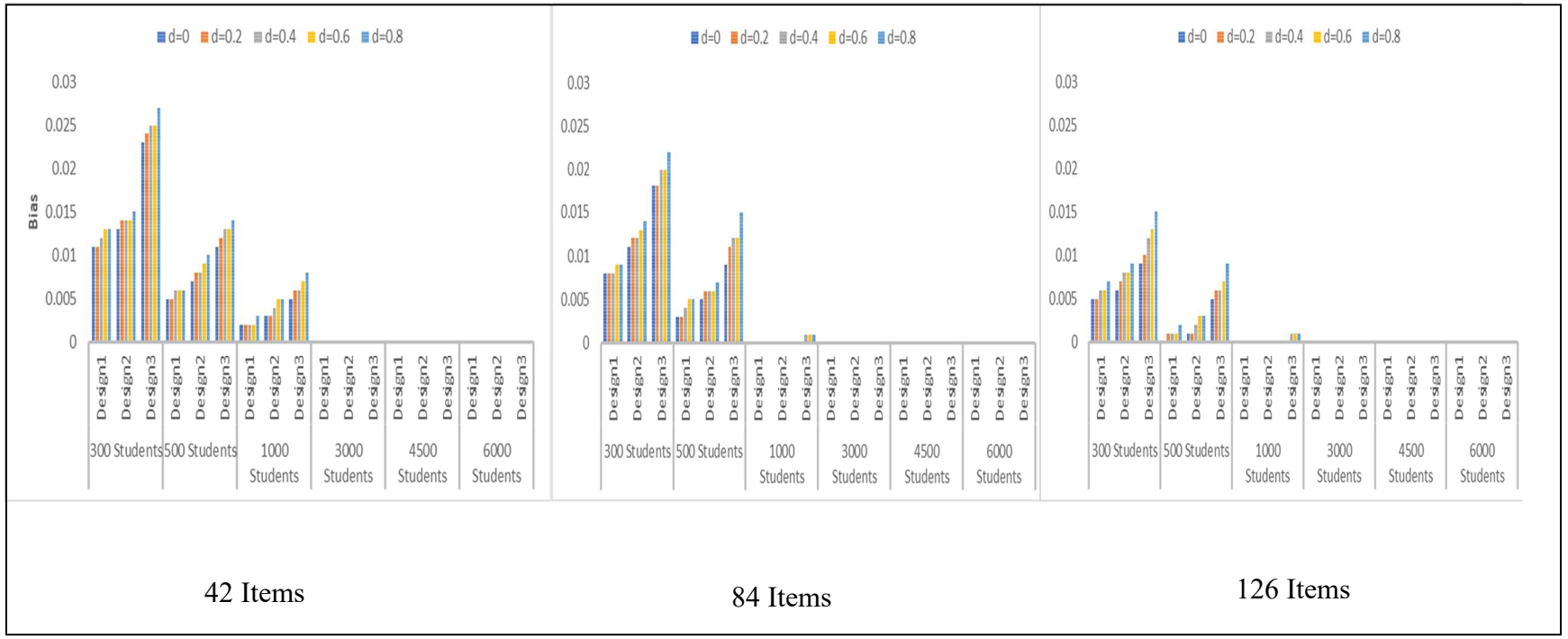
Note. d=0, d=0.2,..., d=0.8 represent the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all the match conditions investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

A.3 RMSE for the recovery of the mean person ability estimate across different distribution match conditions (Case for a test length of 42 items)



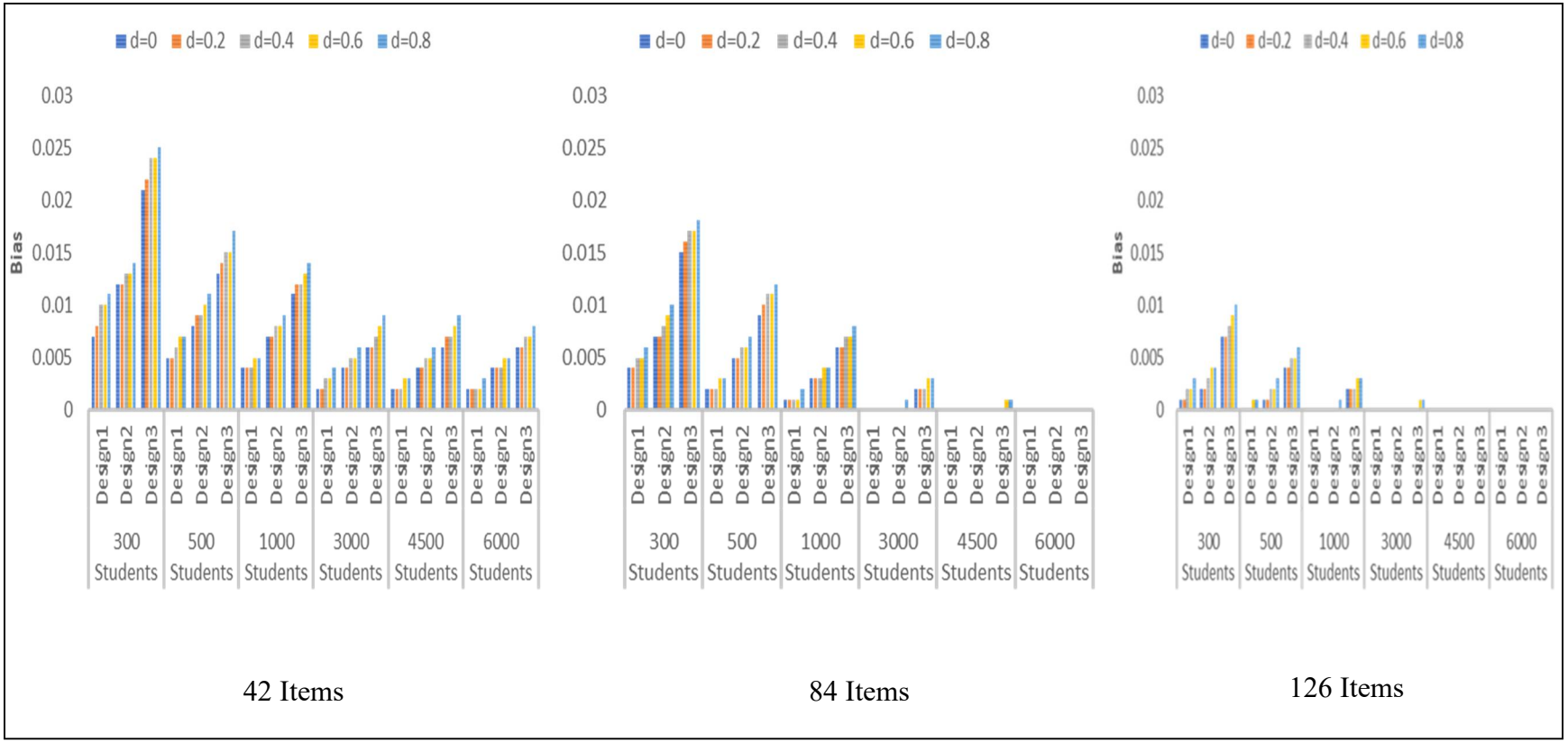
Note. "Distribution Match" represents the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all conditions of distribution match investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "distribution match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively.

A.4 Bias for recovery of variance of θ for different item-person match conditions



Note. θ represents the distribution of person abilities. Also, d=0, d=0.2,..., d=0.8 represent the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all the match conditions investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

A.5 Bias for the recovery of mean item difficulty for different item-person match conditions



Note. d=0, d=0.2,...,d=0.8 represent the degree of match between the distribution of item difficulties and the distribution of person abilities. For the condition d=0, there is perfect match between the distribution of item difficulties and the distribution of person abilities (both distributions having a mean of 0). In all the match conditions investigated, the distribution of item difficulties has a fixed mean of 0. Only the mean for the distribution of person abilities is varied. Thus, for the condition d=0.2, the distribution of person abilities has a mean of 0.2. Similarly, for the "match" conditions d=0.4, d=0.6 and d=0.8, the distribution of person abilities has a mean of 0.4, 0.6 and 0.8 respectively. Also, the multi-matrix designs become sparser moving from Design1 to Design3. Design1 contains 57% missing data, Desing2 contains 71% missing data, while Design3 contains 86% missing data.

Appendix B Program Code

This section of the Appendix reproduces the program code used for the most important algorithms used in this PhD project. Every function listed below is programmed for the R statistical environment (R-3.6.0). The data used is available upon request.

B.1 Generation of booklet designs from real assessment data

The following script generates the various multi-matrix designs investigated from the 2015 VERA-8 Mathematics data for Berlin and Brandenburg. This script produces 1000 unique designs for each of the multi-matrix booklet designs investigated. (Case for the sample size, N =10,000 students).

```
library("TAM")
#Load the VERA_8 Mathematics dataset scored_B1 into the working environment
load("scored_B1.Rdata")
set.seed(15254)
D1 <- scored_B1[sample(10000),sample(42)]

dataset <<- list()
for (i in 1:1000) {
  set.seed(i)
  randomSubset <- D1[sample(nrow(D1), ), sample(ncol(D1), )]
  dataset[[i]] <- randomSubset

}



#1) Creation of the BIB one design
create.BIB.one <- function(X){
  #ordering of item difficulties

  mod1 <- tam(X)
  Diff <- mod1$xsi
  Diff_ordered <- Diff[order(Diff$xsi),]
  items_ordered <- rownames(Diff_ordered)
  items_p_2 <- items_ordered

  #stratify the item difficulties
  easy_BIB <- items_p_2[1:14]
```

```
  average_BIB <- items_p_2[15:28]
  diff_BIB <- items_p_2[29:42]



  #shuffle the items in each stratum
  ##shuffle the items above and convert them to characters
  set.seed(548)
  easy_BIB_2 <- sample(easy_BIB, 14)
  easy_BIB_22 <- as.character(easy_BIB_2)

  set.seed(302)
  average_BIB_2 <- sample(average_BIB, 14)
  average_BIB_22 <- as.character(average_BIB_2)

  set.seed(125)
  diff_BIB_2 <- sample(diff_BIB, 14)
  diff_BIB_22 <- as.character(diff_BIB_2)



  #create item blocks
  BIB_Block_1 <- c(easy_BIB_22[1:2], average_BIB_22[1:2], diff_BIB_22[1:2])


  BIB_Block_2 <- c(easy_BIB_22[3:4], average_BIB_22[3:4], diff_BIB_22[3:4])


  BIB_Block_3 <- c(easy_BIB_22[5:6], average_BIB_22[5:6], diff_BIB_22[5:6])


  BIB_Block_4 <- c(easy_BIB_22[7:8], average_BIB_22[7:8], diff_BIB_22[7:8])


  BIB_Block_5 <- c(easy_BIB_22[9:10], average_BIB_22[9:10],
diff_BIB_22[9:10])


  BIB_Block_6 <- c(easy_BIB_22[11:12], average_BIB_22[11:12],
diff_BIB_22[11:12])
```

```
  BIB_Block_7 <- c(easy_BIB_22[13:14], average_BIB_22[13:14],
diff_BIB_22[13:14])



  scored_B1_p <- X


  #create 7 population blocks which will be administered the different
booklets
  G1 <- scored_B1_p[1:1428,]
  G2 <- scored_B1_p[1429:2856,]
  G3 <- scored_B1_p[2857:4284,]
  G4 <- scored_B1_p[4285:5712,]
  G5 <- scored_B1_p[5713:7140,]
  G6 <- scored_B1_p[7141:8568,]
  G7 <- scored_B1_p[8569:10000,]


  #create the BIBS Sample though with data even for the unneeded blocks
  BIB_Booklet_1 <- G1[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_2 <- G2[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_3 <- G3[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_4 <- G4[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_5 <- G5[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_6 <- G6[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_7 <- G7[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]


  #remove the respective unneeded blocks from the various booklets
  BIB_Booklet_1[c(BIB_Block_3, BIB_Block_5, BIB_Block_6, BIB_Block_7)] <-
NA
  BIB_Booklet_2[c(BIB_Block_1, BIB_Block_4, BIB_Block_6, BIB_Block_7)] <-
NA
  BIB_Booklet_3[c(BIB_Block_1, BIB_Block_2, BIB_Block_5, BIB_Block_7)] <-
NA
```

```
  BIB_Booklet_4[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_6)] <-
NA
  BIB_Booklet_5[c(BIB_Block_2, BIB_Block_3, BIB_Block_4, BIB_Block_7)] <-
NA
  BIB_Booklet_6[c(BIB_Block_1, BIB_Block_3, BIB_Block_4, BIB_Block_5)] <-
NA
  BIB_Booklet_7[c(BIB_Block_2, BIB_Block_4, BIB_Block_5, BIB_Block_6)] <-
NA



  total_BIB_one <<- rbind(BIB_Booklet_1, BIB_Booklet_2, BIB_Booklet_3,
BIB_Booklet_4, BIB_Booklet_5, BIB_Booklet_6, BIB_Booklet_7)


}


#2) Creation of the BIB two design
create.BIB.two <- function(X){
  #ordering of item difficulties


  mod1 <- tam(X)
  Diff <- mod1$xsi
  Diff_ordered <- Diff[order(Diff$xsi),]
  items_ordered <- rownames(Diff_ordered)
  items_p_2 <- items_ordered

  #stratify the item difficulties
  easy_BIB <- items_p_2[1:14]
  average_BIB <- items_p_2[15:28]
  diff_BIB <- items_p_2[29:42]



  #shuffle the items in each stratum
  ##shuffle the items above and convert them to characters
  set.seed(548)
  easy_BIB_2 <- sample(easy_BIB, 14)
  easy_BIB_22 <- as.character(easy_BIB_2)
```

```
  set.seed(302)
  average_BIB_2 <- sample(average_BIB, 14)
  average_BIB_22 <- as.character(average_BIB_2)


  set.seed(125)
  diff_BIB_2 <- sample(diff_BIB, 14)
  diff_BIB_22 <- as.character(diff_BIB_2)



  #create item blocks
  BIB_Block_1 <- c(easy_BIB_22[1:2], average_BIB_22[1:2], diff_BIB_22[1:2])


  BIB_Block_2 <- c(easy_BIB_22[3:4], average_BIB_22[3:4], diff_BIB_22[3:4])


  BIB_Block_3 <- c(easy_BIB_22[5:6], average_BIB_22[5:6], diff_BIB_22[5:6])


  BIB_Block_4 <- c(easy_BIB_22[7:8], average_BIB_22[7:8], diff_BIB_22[7:8])


  BIB_Block_5 <- c(easy_BIB_22[9:10], average_BIB_22[9:10],
diff_BIB_22[9:10])


  BIB_Block_6 <- c(easy_BIB_22[11:12], average_BIB_22[11:12],
diff_BIB_22[11:12])


  BIB_Block_7 <- c(easy_BIB_22[13:14], average_BIB_22[13:14],
diff_BIB_22[13:14])



  scored_B1_p <- X



  #create 7 population blocks which will be administered the different
booklets
  G1 <- scored_B1_p[1:1428,]
  G2 <- scored_B1_p[1429:2856,]
```

```
  G3 <- scored_B1_p[2857:4284,]

  G4 <- scored_B1_p[4285:5712,]

  G5 <- scored_B1_p[5713:7140,]

  G6 <- scored_B1_p[7141:8568,]

  G7 <- scored_B1_p[8569:10000,]


  #create the BIBS Sample though with data even for the unneeded blocks

  BIB_Booklet_1 <- G1[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]

  BIB_Booklet_2 <- G2[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]

  BIB_Booklet_3 <- G3[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]

  BIB_Booklet_4 <- G4[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]

  BIB_Booklet_5 <- G5[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]

  BIB_Booklet_6 <- G6[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]

  BIB_Booklet_7 <- G7[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]


  #remove the respective unneeded blocks from the various booklets

  BIB_Booklet_1[c(BIB_Block_3, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7)] <- NA

  BIB_Booklet_2[c(BIB_Block_1, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7)] <- NA

  BIB_Booklet_3[c(BIB_Block_1, BIB_Block_2, BIB_Block_5, BIB_Block_6,
BIB_Block_7)] <- NA

  BIB_Booklet_4[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_6,
BIB_Block_7)] <- NA

  BIB_Booklet_5[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_7)] <- NA

  BIB_Booklet_6[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5)] <- NA

  BIB_Booklet_7[c(BIB_Block_2, BIB_Block_3, BIB_Block_4, BIB_Block_5,
BIB_Block_6)] <- NA
```

```
  total_BIB_two <<- rbind(BIB_Booklet_1, BIB_Booklet_2, BIB_Booklet_3,
BIB_Booklet_4, BIB_Booklet_5, BIB_Booklet_6, BIB_Booklet_7)


}


#3) Creation of the BIB three design
create.BIB.three <- function(X){
  #ordering of item difficulties


  mod1 <- tam(X)
  Diff <- mod1$xsi
  Diff_ordered <- Diff[order(Diff$xsi),]
  items_ordered <- rownames(Diff_ordered)
  items_p_2 <- items_ordered

  #stratify the item difficulties
  easy_BIB <- items_p_2[1:14]
  average_BIB <- items_p_2[15:28]
  diff_BIB <- items_p_2[29:42]



  #shuffle the items in each stratum
  ##shuffle the items above and convert them to characters
  set.seed(548)
  easy_BIB_2 <- sample(easy_BIB, 14)
  easy_BIB_22 <- as.character(easy_BIB_2)


  set.seed(302)
  average_BIB_2 <- sample(average_BIB, 14)
  average_BIB_22 <- as.character(average_BIB_2)


  set.seed(125)
  diff_BIB_2 <- sample(diff_BIB, 14)
  diff_BIB_22 <- as.character(diff_BIB_2)
```

```
#create item blocks
BIB_Block_1 <- c(easy_BIB_22[1], average_BIB_22[1], diff_BIB_22[1])

BIB_Block_2 <- c(easy_BIB_22[2], average_BIB_22[2], diff_BIB_22[2])

BIB_Block_3 <- c(easy_BIB_22[3], average_BIB_22[3], diff_BIB_22[3])

BIB_Block_4 <- c(easy_BIB_22[4], average_BIB_22[4], diff_BIB_22[4])

BIB_Block_5 <- c(easy_BIB_22[5], average_BIB_22[5], diff_BIB_22[5])

BIB_Block_6 <- c(easy_BIB_22[6], average_BIB_22[6], diff_BIB_22[6])

BIB_Block_7 <- c(easy_BIB_22[7], average_BIB_22[7], diff_BIB_22[7])

BIB_Block_8 <- c(easy_BIB_22[8], average_BIB_22[8], diff_BIB_22[8])

BIB_Block_9 <- c(easy_BIB_22[9], average_BIB_22[9], diff_BIB_22[9])

BIB_Block_10 <- c(easy_BIB_22[10], average_BIB_22[10], diff_BIB_22[10])

BIB_Block_11 <- c(easy_BIB_22[11], average_BIB_22[11], diff_BIB_22[11])

BIB_Block_12 <- c(easy_BIB_22[12], average_BIB_22[12], diff_BIB_22[12])

BIB_Block_13 <- c(easy_BIB_22[13], average_BIB_22[13], diff_BIB_22[13])

BIB_Block_14 <- c(easy_BIB_22[14], average_BIB_22[14], diff_BIB_22[14])


scored_B1_p <- X
```

```
  #create 14 population blocks which will be administered the different
booklets
  G1 <- scored_B1_p[1:714,]
  G2 <- scored_B1_p[715:1428,]
  G3 <- scored_B1_p[1429:2142,]
  G4 <- scored_B1_p[2143:2856,]
  G5 <- scored_B1_p[2857:3570,]
  G6 <- scored_B1_p[3571:4284,]
  G7 <- scored_B1_p[4285:4998,]
  G8 <- scored_B1_p[4999:5712,]
  G9 <- scored_B1_p[5713:6426,]
  G10 <- scored_B1_p[6427:7140,]
  G11 <- scored_B1_p[7141:7854,]
  G12 <- scored_B1_p[7855:8568,]
  G13 <- scored_B1_p[8569:9282,]
  G14 <- scored_B1_p[9283:10000,]


  #create the BIBS Sample though with data even for the unneeded blocks
  BIB_Booklet_1 <- G1[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]
  BIB_Booklet_2 <- G2[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]
  BIB_Booklet_3 <- G3[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]
  BIB_Booklet_4 <- G4[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]
  BIB_Booklet_5 <- G5[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]
  BIB_Booklet_6 <- G6[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
```

```
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB Booklet 7 <- G7[, c(BIB Block 1, BIB Block 2, BIB Block 3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_8 <- G8[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB Block 9, BIB Block 10, BIB Block 11, BIB Block 12, BIB Block 13,
BIB_Block_14)]

  BIB_Booklet_9 <- G9[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB Booklet 10 <- G10[, c(BIB Block 1, BIB Block 2, BIB Block 3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_11 <- G11[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_12 <- G12[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_13 <- G13[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_14 <- G14[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]




  #remove the respective unneeded blocks from the various booklets

  BIB_Booklet_1[c(BIB_Block_3, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_2[c(BIB_Block_1, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA
```

```
  BIB_Booklet_3[c(BIB_Block_1, BIB_Block_2, BIB_Block_5, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_4[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_5[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_6[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_7[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_8[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_9[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_10[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_11[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_10, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_12[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_10, BIB_Block_11, BIB_Block_14)] <- NA

  BIB_Booklet_13[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_10, BIB_Block_11, BIB_Block_12)] <- NA

  BIB_Booklet_14[c(BIB_Block_2, BIB_Block_3, BIB_Block_4, BIB_Block_5,
BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10,
BIB_Block_11, BIB_Block_12, BIB_Block_13)] <- NA


  total_BIB_three <<- rbind(BIB_Booklet_1, BIB_Booklet_2, BIB_Booklet_3,
BIB_Booklet_4, BIB_Booklet_5, BIB_Booklet_6, BIB_Booklet_7, BIB_Booklet_8,
BIB_Booklet_9, BIB_Booklet_10, BIB_Booklet_11, BIB_Booklet_12,
BIB_Booklet_13, BIB_Booklet_14)


}
```

```
#   MMS creator
MMS.creator <-  function(X){
  create.BIB.one(X)
  create.BIB.two(X)
  create.BIB.three(X)

}
```

B.2 Generation of booklet designs from simulated data

The following script shows how the multi-matrix booklet designs were created in the simulation study in chapter six. This script produces 1000 unique designs for each of the multi-matrix booklet designs investigated. It also presents the case for simulating data for 4500 students and for 42 test items; and for match condition where item difficulties have a mean of 0 while person abilities have a mean of 0.8.

```
library("TAM")
library("irtoys")

dataset <<- list()
for(i in 1:1000) {

  set.seed(i)
  items0 <- rnorm(42, 0, 1)
  items1 <- as.data.frame(items0)
  items2 <- sort(items1$items0)

  items <<- cbind(rep(1, 42), items2, rep(0, 42))

  set.seed(i+1000)
  theta <- rnorm(4500, 0.8, 1)

  resp1 <- sim(ip=items, x=theta)
  resp <- as.data.frame(resp1)

  dataset[[i]] <- resp

}

create.BIB.one <- function(X){
  #ordering of item difficulties
    D1 <- X
```

```
mod1 <- tam(D1)
Diff <- mod1$xsi
Diff_ordered <- Diff[order(Diff$xsi),]
items_ordered <- rownames(Diff_ordered)
items_p_2 <- items_ordered

#stratify the item difficulties
easy_BIB <- items_p_2[1:14]
average_BIB <- items_p_2[15:28]
diff_BIB <- items_p_2[29:42]



#shuffle the items in each stratum
##shuffle the items above and convert them to characters
set.seed(548)
easy_BIB_2 <- sample(easy_BIB, 14)
easy_BIB_22 <- as.character(easy_BIB_2)

set.seed(302)
average_BIB_2 <- sample(average_BIB, 14)
average_BIB_22 <- as.character(average_BIB_2)

set.seed(125)
diff_BIB_2 <- sample(diff_BIB, 14)
diff_BIB_22 <- as.character(diff_BIB_2)



#create item blocks
BIB_Block_1 <- c(easy_BIB_22[1:2], average_BIB_22[1:2], diff_BIB_22[1:2])

BIB_Block_2 <- c(easy_BIB_22[3:4], average_BIB_22[3:4], diff_BIB_22[3:4])

BIB_Block_3 <- c(easy_BIB_22[5:6], average_BIB_22[5:6], diff_BIB_22[5:6])
```

```
  BIB_Block_4 <- c(easy_BIB_22[7:8], average_BIB_22[7:8], diff_BIB_22[7:8])


  BIB_Block_5 <- c(easy_BIB_22[9:10], average_BIB_22[9:10],
diff_BIB_22[9:10])


  BIB_Block_6 <- c(easy_BIB_22[11:12], average_BIB_22[11:12],
diff_BIB_22[11:12])


  BIB_Block_7 <- c(easy_BIB_22[13:14], average_BIB_22[13:14],
diff_BIB_22[13:14])




  scored_B1_p <- X


  #create 7 population blocks which will be administered the different
booklets
  G1 <- scored_B1_p[1:642,]
  G2 <- scored_B1_p[643:1284,]
  G3 <- scored_B1_p[1285:1926,]
  G4 <- scored_B1_p[1927:2568,]
  G5 <- scored_B1_p[2569:3210,]
  G6 <- scored_B1_p[3211:3852,]
  G7 <- scored_B1_p[3853:4500,]


  #create the BIBS Sample though with data even for the unneeded blocks
  BIB_Booklet_1 <- G1[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_2 <- G2[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_3 <- G3[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_4 <- G4[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_5 <- G5[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_6 <- G6[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
```

```
  BIB_Booklet_7 <- G7[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]


  #remove the respective unneeded blocks from the various booklets
  BIB_Booklet_1[c(BIB_Block_3, BIB_Block_5, BIB_Block_6, BIB_Block_7)] <-
NA
  BIB_Booklet_2[c(BIB_Block_1, BIB_Block_4, BIB_Block_6, BIB_Block_7)] <-
NA
  BIB_Booklet_3[c(BIB_Block_1, BIB_Block_2, BIB_Block_5, BIB_Block_7)] <-
NA
  BIB_Booklet_4[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_6)] <-
NA
  BIB_Booklet_5[c(BIB_Block_2, BIB_Block_3, BIB_Block_4, BIB_Block_7)] <-
NA
  BIB_Booklet_6[c(BIB_Block_1, BIB_Block_3, BIB_Block_4, BIB_Block_5)] <-
NA
  BIB_Booklet_7[c(BIB_Block_2, BIB_Block_4, BIB_Block_5, BIB_Block_6)] <-
NA



  total_BIB_one <<- rbind(BIB_Booklet_1, BIB_Booklet_2, BIB_Booklet_3,
BIB_Booklet_4, BIB_Booklet_5, BIB_Booklet_6, BIB_Booklet_7)


}


################


create.BIB.two <- function(X){
  #ordering of item difficulties
    D1 <- X
  mod1 <- tam(D1)
  Diff <- mod1$xsi
  Diff_ordered <- Diff[order(Diff$xsi),]
  items_ordered <- rownames(Diff_ordered)
  items_p_2 <- items_ordered


  #stratify the item difficulties
  easy_BIB <- items_p_2[1:14]
```

```
  average_BIB <- items_p_2[15:28]
  diff_BIB <- items_p_2[29:42]



  #shuffle the items in each stratum
  ##shuffle the items above and convert them to characters
  set.seed(548)
  easy_BIB_2 <- sample(easy_BIB, 14)
  easy_BIB_22 <- as.character(easy_BIB_2)

  set.seed(302)
  average_BIB_2 <- sample(average_BIB, 14)
  average_BIB_22 <- as.character(average_BIB_2)

  set.seed(125)
  diff_BIB_2 <- sample(diff_BIB, 14)
  diff_BIB_22 <- as.character(diff_BIB_2)



  #create item blocks
  BIB_Block_1 <- c(easy_BIB_22[1:2], average_BIB_22[1:2], diff_BIB_22[1:2])

  BIB_Block_2 <- c(easy_BIB_22[3:4], average_BIB_22[3:4], diff_BIB_22[3:4])

  BIB_Block_3 <- c(easy_BIB_22[5:6], average_BIB_22[5:6], diff_BIB_22[5:6])

  BIB_Block_4 <- c(easy_BIB_22[7:8], average_BIB_22[7:8], diff_BIB_22[7:8])

  BIB_Block_5 <- c(easy_BIB_22[9:10], average_BIB_22[9:10],
diff_BIB_22[9:10])

  BIB_Block_6 <- c(easy_BIB_22[11:12], average_BIB_22[11:12],
diff_BIB_22[11:12])
```

```
  BIB_Block_7 <- c(easy_BIB_22[13:14], average_BIB_22[13:14],
diff_BIB_22[13:14])



  scored_B1_p <- X



  #create 7 population blocks which will be administered the different
booklets
  G1 <- scored_B1_p[1:642,]
  G2 <- scored_B1_p[643:1284,]
  G3 <- scored_B1_p[1285:1926,]
  G4 <- scored_B1_p[1927:2568,]
  G5 <- scored_B1_p[2569:3210,]
  G6 <- scored_B1_p[3211:3852,]
  G7 <- scored_B1_p[3853:4500,]



  #create the BIBS Sample though with data even for the unneeded blocks
  BIB_Booklet_1 <- G1[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_2 <- G2[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_3 <- G3[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_4 <- G4[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_5 <- G5[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_6 <- G6[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]
  BIB_Booklet_7 <- G7[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7)]



  #remove the respective unneeded blocks from the various booklets
  BIB_Booklet_1[c(BIB_Block_3, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7)] <- NA
```

```
  BIB_Booklet_2[c(BIB_Block_1, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7)] <- NA

  BIB_Booklet_3[c(BIB_Block_1, BIB_Block_2, BIB_Block_5, BIB_Block_6,
BIB_Block_7)] <- NA

  BIB_Booklet_4[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_6,
BIB_Block_7)] <- NA

  BIB_Booklet_5[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_7)] <- NA

  BIB_Booklet_6[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5)] <- NA

  BIB_Booklet_7[c(BIB_Block_2, BIB_Block_3, BIB_Block_4, BIB_Block_5,
BIB_Block_6)] <- NA



  total_BIB_two <<- rbind(BIB_Booklet_1, BIB_Booklet_2, BIB_Booklet_3,
BIB_Booklet_4, BIB_Booklet_5, BIB_Booklet_6, BIB_Booklet_7)


}


##############


#Creation of the BIB three design
create.BIB.three <- function(X){
  #ordering of item difficulties
  D1 <- X
  mod1 <- tam(D1)
  Diff <- mod1$xsi
  Diff_ordered <- Diff[order(Diff$xsi),]
  items_ordered <- rownames(Diff_ordered)
  items_p_2 <- items_ordered


  #stratify the item difficulties
  easy_BIB <- items_p_2[1:14]
  average_BIB <- items_p_2[15:28]
  diff_BIB <- items_p_2[29:42]
```

```
#schuffle the items in each stratum
##schuffle the items above and convert them to characters
set.seed(548)
easy_BIB_2 <- sample(easy_BIB, 14)
easy_BIB_22 <- as.character(easy_BIB_2)

set.seed(302)
average_BIB_2 <- sample(average_BIB, 14)
average_BIB_22 <- as.character(average_BIB_2)

set.seed(125)
diff_BIB_2 <- sample(diff_BIB, 14)
diff_BIB_22 <- as.character(diff_BIB_2)


#create item blocks
BIB_Block_1 <- c(easy_BIB_22[1], average_BIB_22[1], diff_BIB_22[1])

BIB_Block_2 <- c(easy_BIB_22[2], average_BIB_22[2], diff_BIB_22[2])

BIB_Block_3 <- c(easy_BIB_22[3], average_BIB_22[3], diff_BIB_22[3])

BIB_Block_4 <- c(easy_BIB_22[4], average_BIB_22[4], diff_BIB_22[4])

BIB_Block_5 <- c(easy_BIB_22[5], average_BIB_22[5], diff_BIB_22[5])

BIB_Block_6 <- c(easy_BIB_22[6], average_BIB_22[6], diff_BIB_22[6])

BIB_Block_7 <- c(easy_BIB_22[7], average_BIB_22[7], diff_BIB_22[7])

BIB_Block_8 <- c(easy_BIB_22[8], average_BIB_22[8], diff_BIB_22[8])

BIB_Block_9 <- c(easy_BIB_22[9], average_BIB_22[9], diff_BIB_22[9])
```

```
  BIB_Block_10 <- c(easy_BIB_22[10], average_BIB_22[10], diff_BIB_22[10])

  BIB_Block_11 <- c(easy_BIB_22[11], average_BIB_22[11], diff_BIB_22[11])

  BIB_Block_12 <- c(easy_BIB_22[12], average_BIB_22[12], diff_BIB_22[12])

  BIB_Block_13 <- c(easy_BIB_22[13], average_BIB_22[13], diff_BIB_22[13])

  BIB_Block_14 <- c(easy_BIB_22[14], average_BIB_22[14], diff_BIB_22[14])



  scored_B1_p <- X



  #create 14 population blocks which will be administered the different
booklets
  G1 <- scored_B1_p[1:321,]
  G2 <- scored_B1_p[322:642,]
  G3 <- scored_B1_p[643:963,]
  G4 <- scored_B1_p[964:1284,]
  G5 <- scored_B1_p[1285:1605,]
  G6 <- scored_B1_p[1606:1926,]
  G7 <- scored_B1_p[1927:2247,]
  G8 <- scored_B1_p[2248:2568,]
  G9 <- scored_B1_p[2569:2889,]
  G10 <- scored_B1_p[2890:3210,]
  G11 <- scored_B1_p[3211:3531,]
  G12 <- scored_B1_p[3531:3852,]
  G13 <- scored_B1_p[3853:4173,]
  G14 <- scored_B1_p[4174:4500,]


  #create the BIBS Sample though with data even for the unneeded blocks
  BIB_Booklet_1 <- G1[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]
```

```
  BIB_Booklet_2 <- G2[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_3 <- G3[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_4 <- G4[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_5 <- G5[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_6 <- G6[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_7 <- G7[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_8 <- G8[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_9 <- G9[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_10 <- G10[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_11 <- G11[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_12 <- G12[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]

  BIB_Booklet_13 <- G13[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]
```

```
  BIB_Booklet_14 <- G14[, c(BIB_Block_1, BIB_Block_2, BIB_Block_3,
BIB_Block_4, BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8,
BIB_Block_9, BIB_Block_10, BIB_Block_11, BIB_Block_12, BIB_Block_13,
BIB_Block_14)]



  #remove the respective unneeded blocks from the various booklets

  BIB_Booklet_1[c(BIB_Block_3, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_2[c(BIB_Block_1, BIB_Block_4, BIB_Block_5, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_3[c(BIB_Block_1, BIB_Block_2, BIB_Block_5, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_4[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_6,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_5[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_6[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_8, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_7[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_9, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_8[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_10, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_9[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_11,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_10[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_12, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_11[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_10, BIB_Block_13, BIB_Block_14)] <- NA

  BIB_Booklet_12[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_10, BIB_Block_11, BIB_Block_14)] <- NA

  BIB_Booklet_13[c(BIB_Block_1, BIB_Block_2, BIB_Block_3, BIB_Block_4,
BIB_Block_5, BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9,
BIB_Block_10, BIB_Block_11, BIB_Block_12)] <- NA
```

```
  BIB_Booklet_14[c(BIB_Block_2, BIB_Block_3, BIB_Block_4, BIB_Block_5,
BIB_Block_6, BIB_Block_7, BIB_Block_8, BIB_Block_9, BIB_Block_10,
BIB_Block_11, BIB_Block_12, BIB_Block_13)] <- NA


  total_BIB_three <<- rbind(BIB_Booklet_1, BIB_Booklet_2, BIB_Booklet_3,
BIB_Booklet_4, BIB_Booklet_5, BIB_Booklet_6, BIB_Booklet_7, BIB_Booklet_8,
BIB_Booklet_9, BIB_Booklet_10, BIB_Booklet_11, BIB_Booklet_12,
BIB_Booklet_13, BIB_Booklet_14)


}


#########
#   MMS creator
MMS.creator <-  function(X){
  create.BIB.one(X)
  create.BIB.two(X)
  create.BIB.three(X)
  MMS.zero.sim <<- X[1:4500, ]
  MMS.one.sim <<- total_BIB_one[, c(colnames(MMS.zero.sim))]
  MMS.two.sim <<- total_BIB_two[, c(colnames(MMS.zero.sim))]
  MMS.three.sim <<- total_BIB_three[, c(colnames(MMS.zero.sim))]


}


MMS.creator(resp)
```

Erklärung

Hiermit versichere ich, dass ich die Dissertation „*An Examination of Parameter Recovery using Different Multiple Matrix Booklet Designs*“ selbstständig verfasst habe. Alle Hilfsmittel, die ich verwendet habe, sind angegeben. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, im Januar, 2020

_____

Anta Akuro

Curriculum vitae

The Curriculum vitae is not included in the Online version due to data protection reasons.