International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 647–659
https://doi.org/10.21449/ijate.1440316
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

A practical guide to item bank calibration with multiple matrix sampling

Eren Can Aybek[1*], **Serkan Arıkan**[2], **Güneş Ertaş**[2]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Türkiye
[2]Bogazici University, Faculty of Education, Department of Mathematics and Science Education, Türkiye

ARTICLE HISTORY
Received: Feb. 20, 2024
Accepted: Aug. 12, 2024

Keywords:
Multiple matrix sampling,
Item bank development,
Item response theory.

Abstract: When it is required to estimate item parameters of a large item bank, Multiple Matrix Sampling (MMS) design provides an efficient way while minimizing the test burden on students. The current study exemplifies how to calibrate a large item pool using MMS design for various purposes, such as developing a CAT administration. The purpose of the current study is to explain and provide an example of how to use MMS design for item bank calibration. Two functions of **mirt** package, `mirt()` and `multipleGroup()` were compared using real data. The results of the present study showed that the standard `mirt()` function is more practical and makes more precise estimations compared to the `multipleGroup()` function.

1. INTRODUCTION

Multiple matrix sampling, also known as rotated booklet design or matrix sampling, is a technique where different participants answer different item blocks to reduce the number of items that each examinee answers while ensuring content coverage. This design is based on the idea of dividing a large item pool into blocks of items and administering different but linked booklets to examinees. Therefore, the so-called "item sampling" makes it possible to administer a large set of items (Lord, 1962). The rotation of the items or blocks across the booklets allows us to obtain a reliable and valid measurement of the examinees' abilities as a group and accurate item parameters while reducing the burden of excessive testing. This design is commonly used in international large-scale assessments (ILSAs). The utilization of rotated booklet designs has become increasingly popular in ILSAs, serving as an effective means of gathering population achievement level estimations from a large number of individuals through the use of large item pools. Overall, Multiple Matrix Sampling (MMS) (Lord, 1962; Shoemaker, 1973) allows for calibrating large item pools while minimizing the test burden on students.

The item sampling is termed as the rotated booklet design in large-scale assessments (Rutkowski et al., 2010) or multiple matrix sampling (OECD, 2023). This design is used not only in ILSAs, but in any large-scale assessment that intends to calibrate a large item pool, such as when building an item bank in computerized adaptive testing. As stated by Shoemaker (1973), when the item pool is substantial, the MMS design provides a practical advantage for

*CONTACT: Eren Can AYBEK ✉ erencan@aybek.net Pamukkale University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456

estimating the parameters of the items. Also, this design reduces overall testing time and cost for assessment by reducing testing time per examinee and allowing for a more efficient use of resources (Shoemaker, 1973). Overall, Shoemaker (1973) listed the advantages of MMS as follows: MMS reduces the standard error of the estimate and makes it possible to test a large number of items. Also, as participants answer some parts of the items, testing time is reduced.

Thus, when the purpose is to estimate the proficiency distribution of a population, estimate the person parameters, or estimate item parameters using a large item bank, MMS design provides an efficient way to achieve these goals. Integrating IRT and MMS design allows comparable person or item parameters as IRT can estimate these parameters on a common scale. When estimating population parameters, the latent regression IRT model that utilizes item responses and covariates is a widely used model. In this approach, the multiple imputation technique (Rubin, 1987) is used to estimate the plausible values based on the posterior distributions. When the aim is to estimate person parameters, more items per person are needed to increase the measurement precision of individuals, whereas when the aim is to estimate population parameters, increasing the precision for the population is vital (Gonzales & Rutkowski, 2010). When estimating item parameters, various booklet designs are used. These designs are explained in the following section.

1.1. Rotated Booklet Design Types

The requirement to give subtests of items to examinees has prompted the development of various booklet designs. The decision for the specific design is given based on the purpose of the test and the applicability of the design. For computer-based linear tests or paper-based tests, the design needs to be established before finalizing the test booklets. In computerized adaptive tests or multi-stage tests, the items or blocks of items to be administered to examinees are decided based on some algorithms (Gonzales & Rutkowski, 2010).

Gonzales and Rutkowski (2010) categorized booklet designs into complete and incomplete designs. Complete booklet designs are those in which all items or blocks are presented in each form, resulting in all items being answered by all examinees, either in the same order or the rotated order. In complete design, multiple forms can be used by rotating the positions of the items to control the position effect. On the other hand, incomplete booklet designs include booklets that contain a subset of items or blocks. Thus, each examinee answers a subset of all items in the latter one.

Booklet designs are also categorized as balanced and unbalanced designs (Gonzales & Rutkowski, 2010). In a balanced design, every item or block is rotated to appear an equal number of times in each form, whereas in an unbalanced design, some items or blocks rotate, but others generally appear only one time. Balanced booklet designs could control the order effect by counterbalancing.

The balanced incomplete block design (BIBD) was proposed by Lord (1965), in which each subset of items or blocks rotates to appear an equal number of times; therefore, the BIBD balances the position of each item. Table 1 shows one example of a BIBD in which there are a total of 10 items/blocks in the item bank, each student answers five items/blocks, and each item/block appears an equal number of times. On the condition of a large number of items, Shoemaker (1973) investigated the effectiveness of a Partially Balanced Incomplete Block design (PBIBD) compared to a BIBD, finding that the PBIBD could accurately reproduce known means across various conditions. In the PBIBD, each cluster appears a set number of times but does not appear with every other cluster (Rutkowski et al., 2013). A variation of the PIBD was used in TIMSS 2011 and PIRLS 2011.

Table 1. *An example of a balanced incomplete block design.*

Booklet	item1/ block1	item2/ block2	item3/ block3	item4/ block4	item5/ block5	item6/ block6	item7/ block7	item8/ block8	item9/ block9	item10/ block10
1	X	X	X	X	X					
2		X	X	X	X	X				
3			X	X	X	X	X			
4				X	X	X	X	X		
5					X	X	X	X	X	
6						X	X	X	X	X
7	X						X	X	X	X
8	X	X						X	X	X
9	X	X	X						X	X
10	X	X	X	X						X

Table 2 shows one example of an unbalanced incomplete block design (UIBD) in which there are a total of 10 items/blocks in the item bank; each student answers four items/blocks. Items/blocks appear an unequal number of times. Both designs provide links across booklets to calibrate items on the same scale. One of the widely used examples of the BIBD, the BIB7 or Youden squares design, has seven rotated blocks, as shown in Table 3 (Gonzales & Rutkowski, 2010). All blocks are arranged to show up an equal number of times. NAEP, PISA, and TIMSS use designs originated from the BIB7.

Table 2. *An example of an unbalanced incomplete block design.*

Booklet	item1	item2	item3	item4	item5	item6	item7	item8	item9	item10
1	X	X							X	X
2			X	X					X	X
3					X	X			X	X
4							X	X	X	X

Table 3. *BIB7 or Youden squares design.*

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

One commonly used UIBD includes a common part (anchor) and varying blocks. Table 4 depicts an example of such a UIBD, where there is one common block (A) and rotating blocks (B to G). Another version of a UIBD features rotating common parts and non-common parts that appear only once, as depicted in Table 5. In the example given in Table 5, booklet 1 and booklet 2 are linked to each other with C2; booklet 2 and booklet 3 are linked to each other with C3, and so on. For instance, having an item pool of 90 items, the nonrotating part (such as A) might have 10 items, whereas rotating anchors might have 5+5=10 items (such as C1 and C2). Therefore, 90 items could be calibrated while each student answers 20 items in a one-lesson duration.

Table 4. *An example of an unbalanced incomplete block design.*

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

Table 5. *An example of an unbalanced incomplete block design.*

Booklet	Blocks		
1	A	C1	C2
2	B	C2	C3
3	C	C3	C4
4	D	C4	C5
5	E	C5	C6
6	F	C6	C1

1.2. Procedural Issues in MMS

Shoemaker (1973) described some procedural issues regarding the application of MMS. The process of MMS consists of three steps: (a) creating booklets with related items or blocks, (b) administering each booklet to selected examinees, and (c) calibrating item and person parameters. Following these steps raises a number of issues to consider when developing the design. For instance, how many subtests will be created? How many test takers are required per booklet? Which is preferable: making fewer booklets with more items in each, or more booklets with fewer items in each? For a more detailed discussion, please visit the book of Shoemaker.

1.3. MMS Designs in Large-Scale Assessments

To minimize student burden and estimate population parameters, large-scale assessment programs (e.g., PISA, NAEP, PIRLS, and TIMSS) use the MMS design as they have wide content coverage. As the purpose of these large-scale assessments is to make inferences based on the population, individual scores are not provided to participants. Focusing on population parameters instead of sample parameters allows one to use the most appropriate MMS design based on specific purposes (Gonzales & Rutkowski, 2010).

In PISA 2021, for questionnaire sections, a within-construct matrix sampling design was used. In this design, questions rotate within constructs instead of between constructs. Thus, a student answers different subsets of questions for each construct. In PISA 2018 field trial design, many testlets were used to eliminate the item order effect, and then, students were randomly assigned to these testlets (OECD, 2020). PISA also links their assessments to the one that preceded it by anchor booklets.

TIMSS 2023 administration used a group adaptive assessment design while maintaining the 14-block TIMSS design (Table 6). The booklets were composed of difficult (D), medium (M), and easy (E) items. Seven of the fourteen booklets were created with difficult or medium blocks, whereas the other seven were created with medium or easy blocks. The booklets are linked via common blocks. 70% of the students in high-achieving countries were randomly assigned to more difficult booklets and rest were assigned to the easy booklets (30%); for middle-level countries, these percentages were 50% and 50%; and for low-achieving countries, 30% of the students were randomly assigned to more difficult booklets, and the rest were assigned to the easy booklets (70%). The idea is to better match assessment difficulty with student ability in each country (Yin & Foy, 2021).

Table 6. *TIMSS booklet design.*

Booklets		Blocks			
More Difficult Booklets	1	SM1	SD1	MM1	MD1
	2	MD2	MD3	SD2	SD3
	3	SM2	SD2	MM2	MD2
	4	MD5	MD1	SD5	SD1
	5	SM3	SD3	MM3	MD3
	6	MM4	MD4	SM4	SD4
	7	SD4	SD5	MD4	MD5
Less Difficult Booklets	8	ME1	MM1	SE1	SM1
	9	SE1	SE2	ME1	ME2
	10	ME2	MM2	SE2	SM2
	11	SE3	SE5	ME3	ME5
	12	ME3	MM3	SE3	SM3
	13	SE4	SM4	ME4	MM4
	14	ME5	ME4	SE5	SE4

First M: Mathematics; Second M: Medium; S: Science; D: Difficult; E: Easy

1.4. Studies Based on MMS Designs

MMS designs are used to estimate the proficiency distribution of a population, person parameters, or item parameters utilizing a large item bank. In international large scale assessments, the main purpose of using MMS designs is to estimate population parameters. NAEP, TIMSS, and PISA use MMS design to control the item exposure rate and to ensure that an adequate number of items are presented to each individual for estimating population-level achievement (Rutkowski, 2014). Also, another benefit of using a rotated booklet design is minimizing student burden.

Several studies were conducted to compare different designs using Large Scale Assessment data (e.g., PISA). With a focus on investigating missing data imputation and plausible value generation methodologies, Kaplan and Su (2016) conducted studies to compare three distinct designs: the two-form design, the three-form design, and the PBIBD (partially balanced incomplete block matrix sampling design), utilizing data from the PISA 2012. For a similar purpose, Adam et al. (2013) developed and compared two-form MMS designs using data from the PISA 2006. They have also exemplified the use of MMS designs for questionnaires in their study.

Some studies consider estimating item parameters and population-level parameters for questionnaires. Munger and Loyd (1988) showed that the MMS procedure could be used for the mail survey questionnaires. They reported that the response rate was higher in item-sampled questionnaires. When there are many items in a questionnaire, and the purpose is to estimate item parameters, multiple matrix sampling could be used to minimize the participant burden. In her dissertation, Yan Zhou (2021) conducted a simulation study to develop and compare MMS designs, utilizing non-overlapping short blocks to divide a lengthy context questionnaire (CQ).

Simulation studies provide valuable information about different designs and methods. Gressard and Loyd (1991) conducted a Monte Carlo simulation study to examine how item sampling through item stratification influences parameter estimation when utilizing multiple matrix sampling with achievement data. Gonzales and Rutkowski (2010) compared various designs based on a simulation study. They focus on the effects of various designs on estimating person ability estimates and item parameters and discuss key issues for developing a booklet design. They point out that test developers should find a balanced model for their data since different results would be obtained for the real data.

1.5. Present Study

MMS designs are used when a large set of items is required to measure a construct to minimize burden on participants. Like computerized adaptive testing, large scale assessments require a large and calibrated item bank; therefore, the use of rotated booklet design offers advantages in estimating item parameters and developing the item bank. While MMS designs are useful for covering a broad content, minimizing student burden and testing time, and facilitating the estimation of population parameters, estimating item parameters on a common scale requires advanced item analysis techniques. However, the majority of MMS studies focus on estimating student parameters with various designs. Despite the growing number of studies requiring a calibrated large item pool, there is a dearth of literature offering practical guidance on how to estimate item parameters utilizing MMS designs in real datasets. Thus, the purpose of the current study is to explain and provide an example of how to calibrate a large item bank that is given to students with an MMS design. In the current study, it is exemplified how a real item pool, including 540 math items at the fourth-grade level can be calibrated via UIBD.

2. METHOD

2.1. Participants

The current study makes use of items and data from a project that aims to develop a CAT system for fourth graders. In the field test phase, 3108 students- 66% of public schools and 34% of private schools-participated in order to calibrate an item bank including 540 mathematics items. A total of twelve public schools and twenty-three private schools participated in the current research. The schools and the students volunteered to attend the study.

2.2. Instrument

To create a computerized adaptive test system, first, a large item pool of fourth-grade mathematics items, 540 items, was developed. These items were developed based on TIMSS assessment framework where items were planned to measure three types of cognitive dimensions: knowing, applying and reasoning (Mullis et al., 2021). Due to the hierarchical nature of TIMSS taxonomy, knowing items are supposed to be simpler than applying items, whereas reasoning items are the most cognitively demanding. To enable simultaneous calibration of these 540 items, they were placed into 36 booklets, each containing 20 items (see Table 7). Items were placed accordingly to create parallel booklets in terms of content and cognitive dimensions, and applying items were mainly placed to anchor items as applying items are suitable to the majority of the students. This procedure has been done by measurement specialists and math educators according to the test blueprint. Using blocks by grouping items was also useful to maintain the similarity of the item contexts for each booklet. Otherwise, participants' scores could be affected by unequal context distribution, and this situation might create construct-irrelevant variance (Gonzales & Rutkowski, 2010).

The testing time is one of the most significant limitations in actual data collection. Considering that classes often run 40 or 50 minutes, 20 items per student would be considered sufficient. Thus, a UIBD was selected in order to calibrate 540 items while administering the minimum item per student. Complete booklet designs were not selected as they required 540 items to be given to each pupil. Furthermore, the BIBD were not preferred since they necessitated using an equal quantity of each item, which meant making more booklets. For instance, a BIBD with 20 items per a booklet will result in 540 booklets; a very large sample size is needed to calibrate that many booklets. Therefore, to have a minimum number of booklets, a UIBD was selected. In the UIBD, similar to the one in Table 5, 540 items could be calibrated using 36 booklets. In the current study design, the first blocks, like block As, had 10 items, and the anchor blocks, block Bs and block Cs, each had five items. Therefore, we end up with a total of 20 items per booklet and 36 booklets. Booklet 1 is linked to booklet 2 via B1 and to booklet 36 via C18; booklet 2 is linked to booklet 1 via B1 and to booklet 3 via C1, and so on. The total quantity of

booklets will differ based on the number of items in each block; for instance, fewer booklets will be produced overall if there are fewer items in the anchor blocks and more items in the initial blocks. But since fewer items in anchor blocks could raise the standard error, a substantial number of items are needed in anchor blocks.

Table 7. *Multiple Matrix Design of the current study.*

	Unique Items Blocks A (36 Blocks; 10 items each)	Anchor Item Blocks B (18 Blocks; 5 items each)	Anchor Item Blocks C (18 Blocks; 5 items each)
Booklet 1	Block A1 (items 1-10)	Block B1 (items 361-365)	Block C18 (items 536-540)
Booklet 2	Block A2 (items 11-20)	Block B1 (items 361-365)	Block C1 (items 451-455)
Booklet 3	Block A3 (items 21-30)	Block B2 (items 366-370)	Block C1 (items 451-455)
Booklet 4	Block A4 (items 31-40)	Block B2 (items 366-370)	Block C2 (items 456-460)
Booklet 5	Block A5 (items 41-50)	Block B3 (items 371-375)	Block C2 (items 456-460)
Booklet 6	Block A6 (items 51-60)	Block B3 (items 371-375)	Block C3 (items 461-465)
.	.	.	.
.	.	.	.
.	.	.	.
Booklet 33	Block A33 (items 321-330)	Block B17 (items 441- 445)	Block C16 (items 526-530)
Booklet 34	Block A34 (items 331-340)	Block B17 (items 441- 445)	Block C17 (items 531-535)
Booklet 35	Block A35 (items 341-350)	Block B18 (items 446- 450)	Block C17 (items 531-535)
Booklet 36	Block A36 (items 351-360)	Block B18 (items 446- 450)	Block C18 (items 536-540)

2.3. Data Analysis

Student data was collected on the Concerto Platform as a long data format (examinees in rows and variables in columns). Data cleaning and preparations were handled using R (R Core Team, 2023) and the *dplyr* package (Wickham et.al., 2023). Following the administration of the booklets, four items were removed from the dataset as two items had zero variances, and the other two items had a printing error. Then the local independence assumption was checked by using Yen's *Q3* statistic with a 0.20 cut-off criterion (Chen & Thissen, 1997). According to Yen's *Q3* statistics, 23 items that violate local independence assumption were eliminated.

Items were calibrated with the *mirt* package (Chalmers, 2012) using `mirt()` and `multipleGroup()` functions. We refer to the method that uses `mirt()` function as the standard method and `multipleGroup()` function as the multiple group method. The standard method is used for IRT item calibrations according to dichotomous and polytomous IRT models. On the other hand, the multiple group method is utilized for vertical scaling (particular items answered by only one group while both groups answered common anchor items) in addition to its major applications, such as detecting differential item functioning (DIF) and differential test functioning (DTF). It divides the data into subsets, applies the conventional procedure to each subset independently, and then aggregates the outcomes. During this process, multiple group method allows the user to constrain some parameters to be equal (e.g., anchoring). On the other hand, the standard method uses the entire dataset, assigns plausible values to missing data, and then makes the calibrations (Chalmers, 2023).

For the `multipleGroup()` function, booklets were used as the grouping variable. However, because of the `multipleGroup()` function's massive processing power needs, it is typically necessary to perform the estimations as paired pairs in order to estimate the standard errors of item parameter estimates. That's why we run the `multipleGroup()` function for paired booklets: booklet 1 and booklet 2; booklet 2 and booklet 3; booklet 3 and booklet 4, and so on.

Despite the enormous overall number of students in the current study, there were around 90 pupils per booklet. Therefore, the Rasch model was selected to calibrate the item bank using both methods (O'Neill et.al., 2020). Then, the difficulty (b) parameters and their standard errors

for both methods were compared.

In order to evaluate the consistency of b parameters, the correlation between IRT b parameters and Classical Test Theory (CTT) *p* statistics were estimated. Research showed that under the CTT and IRT frameworks, there is a strong correlation between item difficulty parameters (MacDonald & Paunonen, 2002). The significance of the difference between these correlations obtained from both calibration methods was tested by using Fisher's Z test, and Cohen's *q* statistics for the effect size. The calculations for the Fisher's Z test and Cohen's *q* statistics were handled with the *diffcor* package (Blotner, 2024) in R. The R codes used in the data analysis can be reached through https://github.com/ecaybek/rbd

3. FINDINGS

3.1. Comparison of b Parameters

The item difficulty parameters were calibrated using the Rasch model, and descriptive statistics for the b parameters are presented in Table 8 for the multiple group method and standard method. The results showed that the item bank covered an ability range of -4.66 to 2.90 for the multiple group method and -4.62 to 2.88 for the standard method. The mean of the b parameters for both methods were close to zero and b parameters were normally distributed according to both methods.

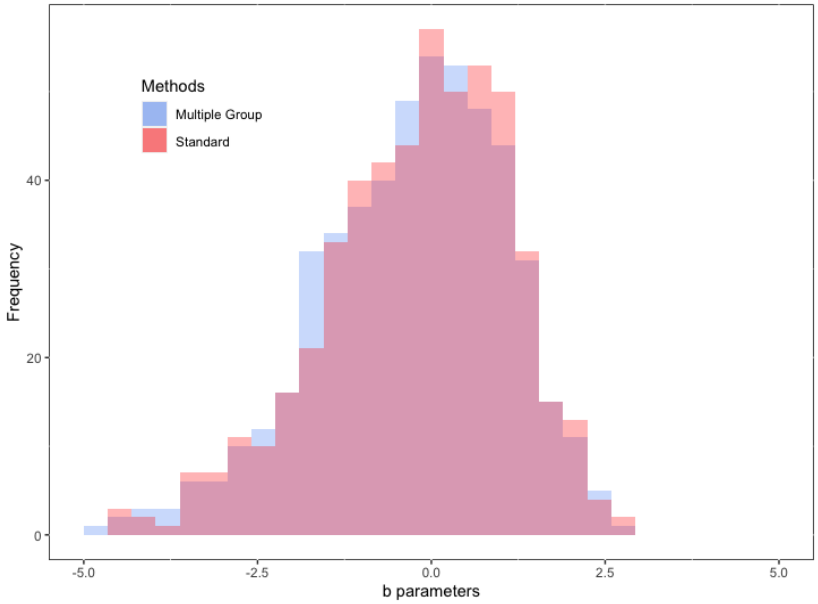
Table 8. *Descriptive statistics of b parameter estimations by two methods.*

Methods	*k*	min	max	mean	median	*s*	skewness	kurtosis
Multiple Group	513	-4.66	2.90	-0.20	-0.13	1.36	-0.52	0.10
Standard	513	-4.62	2.88	-0.21	-0.08	1.35	-0.54	0.17

k: number of items; s: standard deviation

The mean difference of b parameters between the two methods was not significant (t_{1024} = -0.90; p = .37). The distribution of the b parameters for the multiple group and the standard method is presented in Figure 1. As can be seen in Table 8 and Figure 1, according to the estimations from both methods, the item bank had items targeting a very large range of ability levels, especially for very low ability levels (lower than -2) and high ability levels (higher than 2).

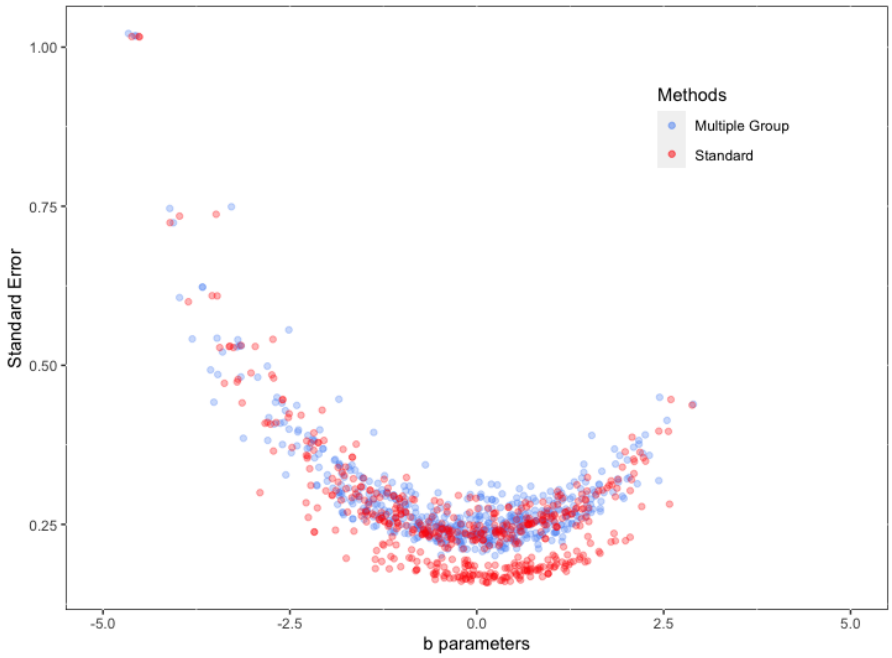
Figure 1. *Distribution of the b parameters of the item bank.*



3.2. Evaluation of Standard Errors

The standard errors of the b parameters estimated by both methods were compared to gain a better understanding of the item parameter estimations (see Figure 2). The standard method tends to estimate b parameters with smaller standard errors than the multiple group method. This discrepancy may be due to the `multipleGroup()` function's enormous processing power requirements.

Figure 2. *Distribution of the SEs of the b parameters of the item bank.*



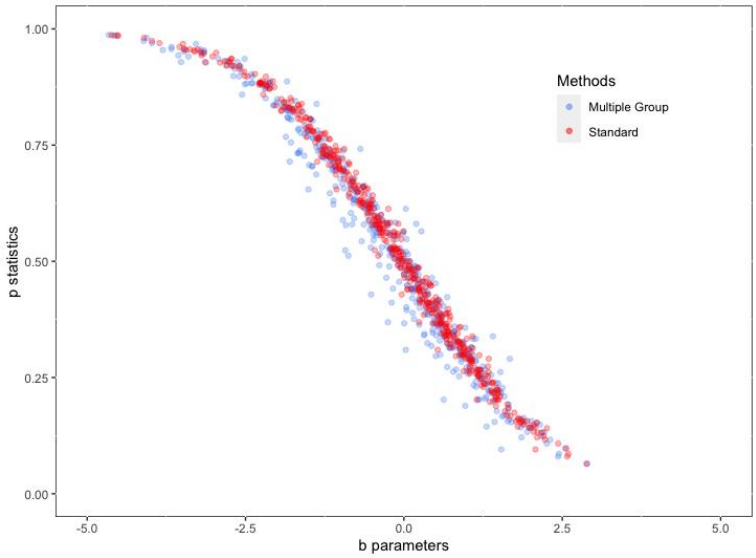
Because of this need, b parameters are estimated by using booklet pairs (Booklet 1 - Booklet 2, Booklet 2 - Booklet 3, and so on) with `multipleGroup()` function. Because the multiple group methodology used data from booklet pairs, while the standard method used the complete dataset, the multiple group method likely estimated the b parameters with higher standard errors due to the smaller dataset size. The U-shaped plot of the standard error occurs due to relatively easy and difficult items having fewer observations for estimating the lower asymptote (Thissen & Wainer, 1982). We believe that the items at the tails have very similar standard errors for both methods.

3.3. Correlation among IRT and CTT Difficulty Parameters

It is also important to evaluate the correlation between IRT b and the CTT *p* statistics. Since CTT *p* statistics and the IRT b parameters are related to the area under the normal distribution curve, this investigation provided us insight into how well the two methods estimated the item parameters. Thus, the scatter plot between the IRT b parameters and the CTT *p* statistics for both methods is shown in Figure 3.

The scatter plot shows that the IRT b parameter estimates from both methods highly correlate with the CTT *p* statistics. On the other hand, the standard method has a stronger relationship with the CTT *p* statistics. While the correlation coefficient between multiple groups and the CTT was found to be -0.972, the correlation coefficient between the standard method and the CTT was found to be -0.981. Fisher's Z test showed that the standard method had a significantly higher correlation with the CTT *p* statistics than the multiple group method ($z = 3.059$; $p < .01$). On the other hand, Cohen's *q* was found to be 0.19, which indicates the size of the difference was small (Cohen, 1988).

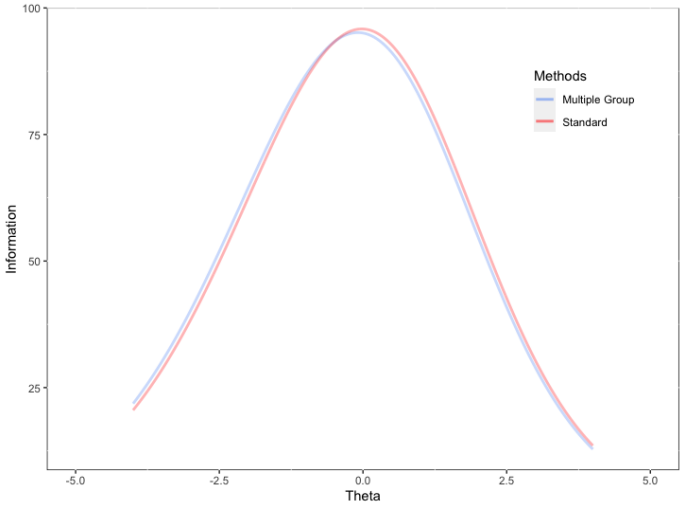
Figure 3. *Scatter plot of the b parameters and the p statistics.*



3.4. Comparison of Test Information Functions

Finally, the test information functions of the item bank were drawn using both methods (Figure 4), and both methods generated very similar test information functions. To sum up all the findings, the standard method has been found more efficient by the manner of computing power and simplicity while there were no significant differences between mean b parameter estimations; the standard method estimates the b parameters with smaller standard error and higher correlation with the CTT *p* statistic.

Figure 4. *Information functions of the item bank.*



4. DISCUSSION and CONCLUSION

The current study aims to exemplify how to calibrate an item bank utilizing MMS design for various purposes, such as developing a CAT administration. Therefore, the current study focuses on why, when, and how to use MMS design. Studies on MMS mostly focus on estimating student parameters, and to the best of our knowledge, estimating item parameters in MMS designs is not prevalent in the literature. Thus, there is a need to demonstrate how to calibrate a large item bank using Multiple Matrix Sampling. Calibrating a large item pool requires deciding on a specific booklet design by considering methodological and practical issues. As Gonzalez and Rutkowksi (2010) stated, in any design, there is a trade-off between what is desired and what is practical based on the purpose of the assessment and existing

resources. More items mean more precision; however, it is more laborious. Integrating the benefits of IRT and MMS, it is more practical and efficient to estimate item parameters of large item pools. Given the constraints of data collection, such as class time of schools and low stake consequences of data collection for participants, it is a kind of must to administer a relatively restricted number of items to students. Depending on the topic, student level and cognitive load, 15 to 20 items may be ideal to administer in a single course time.

In the current study, items (4th grade mathematics) were developed based on the TIMSS Assessment Framework. TIMSS fourth-grade mathematics assessment included three content domains: (1) number, (2) measurement and geometry, (3) data, and three cognitive domains: (1) knowing, (2) applying, (3) reasoning. A substantial number of items within each category should have been administered to enable precise estimation of proficiency distribution (Rutkowski et al., 2013). A total of 540 items were developed in this study. Obviously, it was impossible to administer every item to all examinees. One of the appropriate models to calibrate these items was an unbalanced incomplete booklet design. Thus, in a single lesson period, each student encountered 20 items from all content and cognitive areas.

As simulation studies provided somewhat clean results, using real data from a test provides valuable information and is important for sharing the experience. As Gonzales and Rutkowski (2010) stated, test developers should find a balanced model for their data since different results would be obtained for the real data. Thus, the current study explained the procedures and challenges of calibrating a large item pool using real data.

Each item in the current study was responded to by a varying number of participants due to the design and challenges in reaching out to a big sample. With 36 booklets and 540 items to calibrate, anchor items were answered by approximately 180 students, while non-anchor items were answered by approximately 90 students. As a result, the mean standard error of anchor items was smaller than non-anchor items. In a balanced design, the number of students per item for both anchor and non-anchor items would be similar, resulting in similar standard errors. However, balanced designs will result in more booklets, which require more pupils.

The standard errors of item difficulties were higher for items with extreme difficulties. The estimates of difficulty for items that were very easy and very difficult were less precise compared to the items with medium level difficulty. Gonzalez and Rutkowksi (2010) also reported a similar finding and reported that having more people responding to the items, the precision increases, especially for the extremes. On the one hand, this is a predictable outcome; an item bank for a CAT administration necessitates a huge number of extreme items in order to adequately match student abilities.

The *mirt* package provides very useful tools not only for the conventional item bank development process but also for item bank development under the MMS design. The package includes two functions, `mirt()` and `multipleGroup()`, which are very useful for MMS design. The results of the present study showed that the standard `mirt()` function is more practical and makes more precise estimations when it is compared to the `multipleGroup()` function. It is practical because when `multipleGroup()` function was used with booklet pairs, the estimations took around 42 seconds, while `mirt()` function estimated the item parameters in around 24 seconds. Moreover, the `multipleGroup()` function was incapable of calculating standard errors when 36 booklets were simultaneously included in the analysis. The standard error estimation failed with support not only from the personal computers of the researchers but also from Google Cloud servers. Even though there was no significant difference between the mean of b parameter estimations from both methods, `mirt()` function also estimated the b parameters with less standard error and showed higher correlation with the CTT *p* statistics.

Overall, comparing the multiple group method and standard method, while there were no statistically significant differences between the mean b parameter estimations, the standard method was found to be more efficient in terms of computing power and simplicity. It also estimates b parameters with a smaller standard error and a higher correlation with the CTT *p* statistic.

4.1. Further Suggestions and Limitations

For practical researchers, the standard `mirt()` function is more useful and precise than the `multipleGroup()` function for calibrating item banks with the MMS design. Also, a simulation study can be conducted to compare the bias and RMSE values of the b parameter estimations from both methods. Counterbalancing could also be used to minimize the effect of item order.

One limitation of the current study is that the Rasch model was used to evaluate item discrimination. Due to sample size per booklet, the Rasch model was chosen. A larger sample size per booklet would be better to test the other IRT models. Another limitation is the pairing of booklets when making calibrations via `multipleGroup()` function due to its computational requirements. It would be good to compare the results of this function by running without pairing the booklets.

Acknowledgments

This study was supported by Bogazici University Scientific Research Commission (Project no: BAP-SUP 17002). The preliminary results were presented in 1st Adaptive Test Research National Symposium in 14th - 15th September, 2023 in Bogazici University, Istanbul, Türkiye.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Bogazici University, 84391427-050.01.04-E.9920.

Contribution of Authors

Eren Can Aybek: Research Design, Methodology, Data Collection, Data Analysis, and Writing. **Serkan Arıkan:** Literature Review, Research Design, Methodology, Data Collection, Supervision, Writing and Critical Review. **Güneş Ertaş:** Literature Review, Methodology, Writing, and Data Collection.

Orcid

Eren Can Aybek https://orcid.org/0000-0003-3040-2337
Serkan Arıkan https://orcid.org/0000-0001-9610-5496
Güneş Ertaş https://orcid.org/0000-0001-8785-7768

REFERENCES

Blötner, C. (2024). *Package 'diffcor'*. https://cran.r-project.org/web/packages/diffcor/diffcor.pdf

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433–448). Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R Environment. *Journal of Statistical Software*, *48*, 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R.P. (2023). *Package "mirt"*. https://cran.r-project.org/web/packages/mirt/mirt.pdf

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Gonzalez, E., & Rutkowski, L. (2010). *Principles of Multiple Matrix Booklet Designs and Parameter Recovery in Large-Scale Assessments* (pp. 125–156). IERI.

Gressard, R.P., & Loyd, B.H. (1991). A comparison of item sampling plans in the application of multiple matrix sampling. *Journal of Educational Measurement*, *28*(2), 119–130.

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.

Lord, F.M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*(2), 259–267. https://doi.org/10.1177/001316446202200202

Lord, F.M. (1965). Item sampling in test theory and in research design. *ETS Research Bulletin Series*, *1965*(2), i–39. https://doi.org/10.1002/j.2333-8504.1965.tb00968.x

Macdonald, P., & Paunonen, S.V. (2002). A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921-943. https://doi.org/10.1177/0013164402238082

Munger, G.F., & Loyd, B.H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), 187–191.

OECD. (2020). *PISA 2018 Technical Report-PISA*. OECD Publishing, Paris. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2023). *PISA 2022 Technical Report-PISA*. OECD Publishing, Paris. Retrieved from https://www.oecd.org/pisa/data/pisa2022technicalreport/

O'Neill, T.R., Gregg, J.L., & Peabody, M.R. (2020). Effect of sample size on sommon item equating using the dichotomous rasch model. *Applied Measurement in Education*, *33*(1), 10–23. https://doi.org/10.1080/08957347.2019.1674309

Rubin, D.B. (2009). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education*, *27*(2), 115-132. https://doi.org/10.1080/08957347.2014.880440

Rutkowski, L., Gonzalez, E., Davier, M. von, & Zhou, and Y. (2013). Assessment design for international large-scale assessments. In *Handbook of International Large-Scale Assessment*. Chapman and Hall/CRC.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in Secondary Analysis and Reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

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

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4).

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science*. O'Reilly Media, Inc.

Yin, L., & Foy, P. (2023). *TIMSS 2023 Assessment Design*. In I.V.S. Mullis, M.O. Martin, & M. von Davier (Eds.), TIMSS 2023 Assessment Frameworks. Boston College, TIMSS & PIRLS International Study Center.

Zhou, Y. (2021). *Improving Multiple Matrix Sampling Design for Questionnaires*. Indiana University.