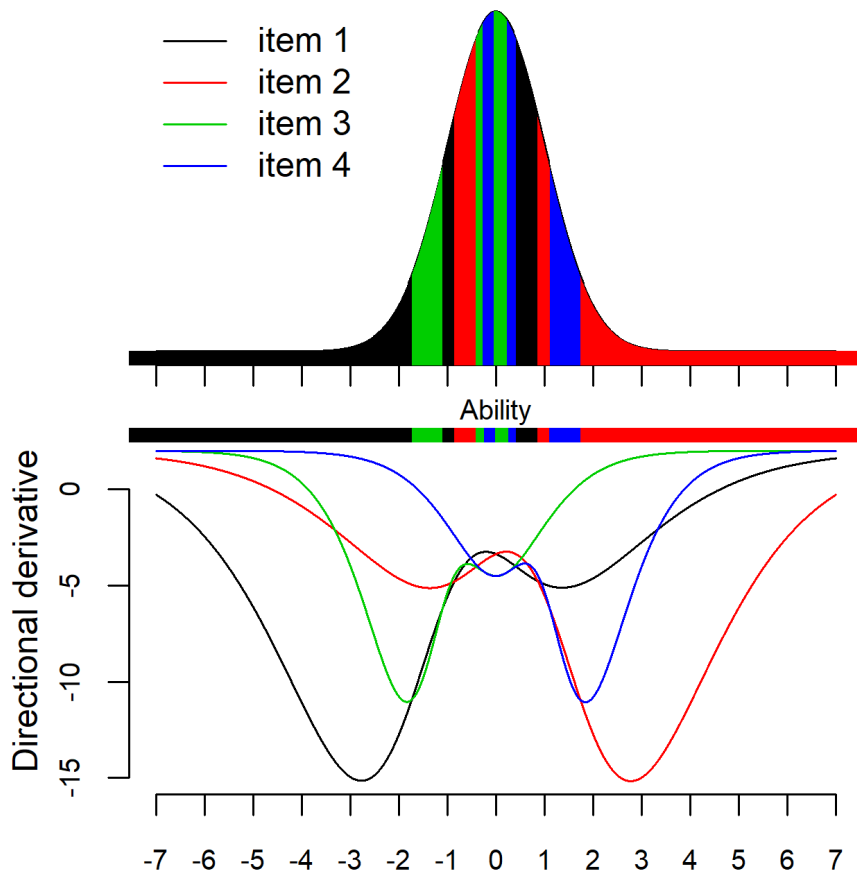


Achievement tests and optimal design for pretesting of questions

Mahmood Ul Hassan



Achievement tests and optimal design for pretesting of questions

Mahmood Ul Hassan

Academic dissertation for the Degree of Doctor of Philosophy in Statistics at Stockholm University to be publicly defended on Friday 15 November 2019 at 10.00 in William-Olssonsalen, Geovetenskapens hus, Svante Arrhenius väg 14, floor 1.

Abstract

Achievement tests are used to measure the students' proficiency in a particular knowledge. Computerized achievement tests (e.g. GRE and SAT) are usually based on questions available in an item bank to measure the proficiency of students. An item bank is a large collection of items with known characteristics (e.g. difficulty). Item banks are continuously updated and revised with new items in place of obsolete, overexposed or flawed items over time. This thesis is devoted to updating and maintaining the item bank with high-quality questions and better estimations of item parameters (item calibration).

The thesis contains four manuscripts. One paper investigates the impact of student ability dimensionality on the estimated parameters and the other three deal with item calibration.

In the first paper, we investigate how the ability dimensionality influences the estimates of the item-parameters. By a case and simulation study, we found that a multidimensional model better discriminates among the students.

The second paper describes a method for optimal item calibration by efficiently selecting the examinees based on their ability levels. We develop an algorithm which selects intervals for the students' ability levels for optimal calibration of the items. We also develop an equivalence theorem for item calibration to verify the optimal design.

The algorithm developed in Paper II becomes complicated with the increase of number of calibrated items. So, in Paper III we develop a new exchange algorithm based on the equivalence theorem developed in Paper II.

Finally, the fourth paper generalizes the exchange algorithm described in Paper III by assuming that the students have multidimensional abilities to answer the questions.

Keywords: *Achievement test, Equivalence theorem, Exchange algorithm, Item calibration, Item response theory model, Optimal experimental design.*

Stockholm 2019

<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-174079>

ISBN 978-91-7797-879-4
ISBN 978-91-7797-880-0

Department of Statistics

Stockholm University, 106 91 Stockholm



ACHIEVEMENT TESTS AND OPTIMAL DESIGN FOR PRETESTING
OF QUESTIONS

Mahmood Ul Hassan



Achievement tests and optimal design for pretesting of questions

Mahmood Ul Hassan

©Mahmood Ul Hassan, Stockholm University 2019

ISBN print 978-91-7797-879-4

ISBN PDF 978-91-7797-880-0

Printed in Sweden by Universitetservice US-AB, Stockholm 2019

To my Parents.

Abstract

Achievement tests are used to measure the students' proficiency in a particular knowledge. Computerized achievement tests (e.g. GRE and SAT) are usually based on questions available in an item bank to measure the proficiency of students. An item bank is a large collection of items with known characteristics (e.g. difficulty). Item banks are continuously updated and revised with new items in place of obsolete, overexposed or flawed items over time. This thesis is devoted to updating and maintaining the item bank with high-quality questions and better estimations of item parameters (item calibration).

The thesis contains four manuscripts. One paper investigates the impact of student ability dimensionality on the estimated parameters and the other three deal with item calibration.

In the first paper, we investigate how the ability dimensionality influences the estimates of the item-parameters. By a case and simulation study, we found that a multidimensional model better discriminates among the students.

The second paper describes a method for optimal item calibration by efficiently selecting the examinees based on their ability levels. We develop an algorithm which selects intervals for the students' ability levels for optimal calibration of the items. We also develop an equivalence theorem for item calibration to verify the optimal design.

The algorithm developed in Paper II becomes complicated with the increase of number of calibrated items. So, in Paper III we develop a new exchange algorithm based on the equivalence theorem developed in Paper II.

Finally, the fourth paper generalizes the exchange algorithm described in Paper III by assuming that the students have multidimensional abilities to answer the questions.

Keywords: Achievement test, Equivalence theorem, Exchange algorithm, Item calibration, Item response theory model, Optimal experimental design.

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: Ul Hassan, M. and Miller, F. (2019). Discrimination with unidimensional and multidimensional item response theory models for educational data.
(Under review)

PAPER II: Ul Hassan, M. and Miller, F. (2019). Optimal item calibration for computerized achievement tests, *Psychometrika*.
DOI: 10.1007/s11336-019-09673-6.

PAPER III: Ul Hassan, M. and Miller, F. (2019). An exchange algorithm for optimal calibration of items in computerized achievement test.
(Manuscript)

PAPER IV: Ul Hassan, M. and Miller, F. (2019). Optimal calibration of items for multidimensional achievement tests.
(Manuscript)

Acknowledgements

First, I would like to thank my supervisor Frank Miller for his guidance, encouragement and dedicating time to me whenever it was needed. He made me push past my limits to accomplish my goals.

I would also like to thank all colleagues who have made the department of Statistics a great work place. I am grateful for the support of all my Ph.D fellows, especially Edgar for lots of discussion about different problems. I am also thankful to Daniel Thorburn and my co-supervisor Ellinor Fackel-Fornuius for valuable suggestions and guideline for my research.

I am eternally thankful to my family for their prayers and support throughout my studies.

Most importantly, I would like to express my deep appreciation to my wife Aniqah for her love and caring attitude and for giving birth to an adorable and wonderful daughter Zoyah Hassan.

Contents

Abstract	i
List of Papers	iii
Acknowledgements	v
1 Introduction	1
1.1 Background	1
1.2 Aims of the thesis	2
1.3 Outline	2
2 Item response theory models	3
2.1 Unidimensional item response theory (UIRT) models	3
2.2 Multidimensional item response theory (MIRT) models	5
3 Optimal design theory for item calibration	7
3.1 Optimal unrestricted design	7
3.2 Example	9
3.3 Optimal restricted design	11
3.4 Efficiency comparison	15
4 Overview of the papers and future research	17
5 Sammanfattning	21
References	23

1. Introduction

1.1 Background

Achievement tests are designed to evaluate the proficiency of the students' knowledge, understanding and skill in a specific field of study or program. Its purpose is to measure the examinees acquired knowledge or developed skill which they have learned in a particular subject or group of subjects. Achievement tests are of high importance among teachers, professional associations and employers.

Computerized achievement tests are growing significantly due to increase of computer use in education and growing interest in distance learning educational programs. Now, many government departments, private organizations and educational institutes are using computerized achievement tests for better evaluation of examinees skills and knowledge in a particular subject or different composite of multiple subjects. The most famous large scale achievement tests, e.g. Graduate Record Examinations (GRE) test, Scholastic Assessment Test (SAT), Test of English as a Foreign Language (TOEFL), Programme for International Student Assessment (PISA) are all computerized tests.

Computerized achievement tests usually use an item bank. An item bank is a large collection of items with known characteristics (e.g. difficulty, discrimination and guessing) to measure the knowledge or skills of the students. For better evaluation of the students' knowledge in achievement tests, we need item banks with updated high-quality questions. After some period of time some questions (items) become obsolete or exposed out among the students (Zheng and Chang, 2017). So, we always need to update the item bank continuously with new items. For this, we use item calibration. Item calibration is a process to estimate characteristics of a set of items for inclusion in an item bank so that we can use these items for future testing. Item calibration is a tool for maintaining and updating the item bank. We use sampling or calibration design which consist of a sample of examinees with known or approximately known abilities for efficient estimation of item parameters.

Item response theory (IRT) models are used to characterize the items in an achievement test. The IRT models describe the relationship between probability of correctly responding to an item and the examinee ability. The details of

IRT models are described in Chapter 2. The appropriate selection of the model is an important task in achievement tests (e.g. GRE and SAT) for calibration of items. If we have an item bank with better characterized items we could better evaluate the students in achievement tests. In the first paper, we have real data of an achievement test from higher education. We implemented different IRT models for data analysis and investigate the appropriateness of models through different approaches. In the other three research papers, we assume an appropriate model for the items and develop methods and algorithms for calibration of items.

1.2 Aims of the thesis

The general purpose of the thesis is to develop methods and algorithms which can be used for optimal item calibration.

1. The first aim is to describe how to select the appropriate IRT model for better characterization of items based on a data set from a real achievement test and to investigate from the case study and a simulation study that dimensionality influences the estimation of parameters.
2. The second aim is to develop a method and algorithms for item calibration in unidimensional and multidimensional cases which utilize the students' ability for better estimation of item characteristics. This is the approach in Paper II-IV.

1.3 Outline

The outline of the thesis is as follows. Chapter 2 introduces the some common and important item response theory models which are used in achievement tests. Chapter 3 describes the optimal design theory for item calibration. The overview of papers with some suggestion for future research constitutes the last chapter.

2. Item response theory models

In this chapter, we briefly give an overview of some important item response theory models used in achievement tests. Modern educational testing is based on a theory known as item response theory (IRT). The IRT models describe the relationship between the examinee's ability to answer the question (latent trait, unobservable attribute or characteristic) and the pattern of item responses. The comprehensive details of IRT models are available in van der Linden (2016). These models have gained much popularity for scoring of tests, item calibration, and measuring abilities. We categorize IRT models into unidimensional and multidimensional models based on the examinees' ability dimension.

2.1 Unidimensional item response theory (UIRT) models

In unidimensional IRT models, we assume that the examinees require a single latent ability or some composite of multiple abilities to answer the item. Unidimensional IRT models are useful when each item is designed to measure single latent ability. The common IRT models for dichotomous responses are the one-parameter logistic (1PL), two-parameter logistic (2PL), three-parameter logistic (3PL) and the four-parameter logistic (4PL) model. The probability of a correct response (denoted by $Y=1$) to item i by examinee with ability θ is mathematically modeled by the following item response functions:

1PL

$$p_i(\theta) = P(Y = 1 | \theta, b_i) = \frac{1}{1 + e^{-(\theta - b_i)}}, \quad (2.1)$$

2PL

$$p_i(\theta) = P(Y = 1 | \theta, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}, \quad (2.2)$$

3PL

$$p_i(\theta) = P(Y = 1 | \theta, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}, \quad (2.3)$$

4PL

$$p_i(\theta) = P(Y = 1 | \theta, a_i, b_i, c_i, g_i) = c_i + \frac{g_i - c_i}{1 + e^{-a_i(\theta - b_i)}}, \quad (2.4)$$

In these equations, $p_i(\theta)$ is a probability of correctly response to item i by examinee with ability θ , $a_i \in (0, \infty)$ is the discrimination (slope) parameter

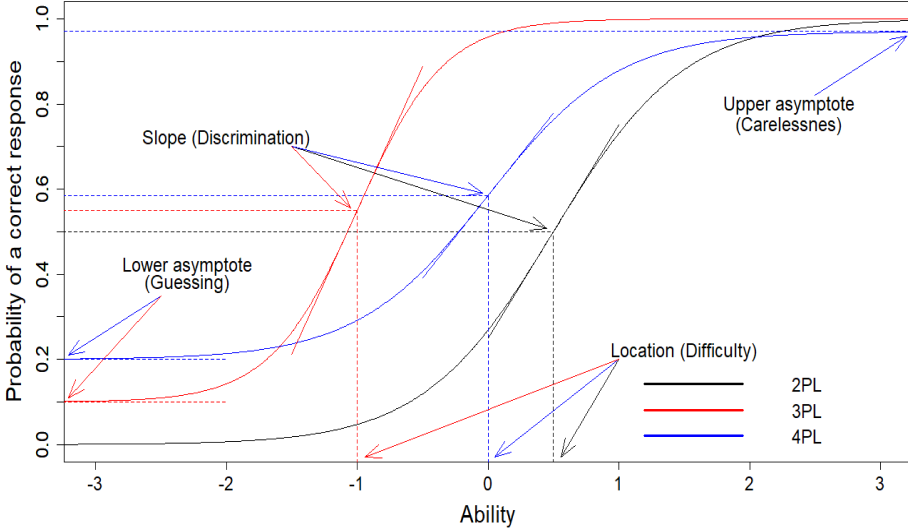


Figure 2.1: The plot shows item characteristic curves of three items following different IRT model.

of item i , $b_i \in \mathbb{R}$ is the difficulty (location) parameter of item i , $c_i \in [0, 1)$ is the guessing (lower/left asymptote) parameter of item i and $g_i \in (0, 1]$ is the upper/right asymptote of item i and θ represents the students ability to answer the question. For practical range of item parameters see Buyske (2005) or Paper III.

We use a polytomous IRT model when the response of an item is scored with more than two levels. For ordered categorical responses, we can use the graded-response (GR) model which was first introduced by Samejima (1969). This model is a generalization of the 2PL model. The probability of the examinee to respond the k^{th} out of r_i categories for the i^{th} item is modeled as

$$\begin{aligned}
 p_{ik}(\theta) &= P(Y = k | \theta; a_i, b_i) = P(Y \geq k | \theta; a_i, b_i) - P(Y \geq k + 1 | \theta; a_i, b_i), \\
 P(Y \geq k | \theta; a_i, b_i) &= \frac{1}{1 + e^{-a_i(\theta - b_{ik})}}, \quad k = 1, 2, \dots, r_i - 1, \quad (2.5) \\
 P(Y \geq 0 | \theta; a_i, b_i) &= 1, \\
 P(Y \geq r_i | \theta; a_i, b_i) &= 0,
 \end{aligned}$$

where $p_{ik}(\theta)$ is a probability of correctly response to k^{th} category of item i by examinee with ability θ , a_i is the discrimination (slope) parameter of item i

and $b_i = (b_{i1}, \dots, b_{i(r_i-1)})$ is a vector of category intercepts for item i .

Other important polytomous IRT models are the partial credit model (Masters, 1982), the rating scale model (Andrich, 1978), the nominal response model (Bock, 1972) and the generalized partial credit model (Muraki, 1992). For a comprehensive introduction of commonly used polytomous IRT models, see van der Linden (2016); van der Linden and Hambleton (2013).

2.2 Multidimensional item response theory (MIRT) models

Multidimensional item response theory (MIRT) models are used to characterize the item when the assumption of unidimensionality is violated. MIRT models are useful when items in an achievement test are designed to measure two or more latent abilities.

We characterize MIRT models as compensatory or non-compensatory models depending on the availability of compensation of high ability on one trait to the lower ability on the other trait. In compensatory MIRT models, examinees with high proficiency on one trait compensate the low proficiency on another trait. However, in non-compensatory MIRT model, this compensation is not available and examinees must be proficient in each ability traits to answer the item correctly. According to Yang (2007), the application of the compensatory models dominate the educational research literature and he lists the following references: Drasgow and Parsons (1983); Kirisci et al. (2001); Reckase (1979); Reckase and McKinley (1991); Way et al. (1988); Yen (1984). The compensatory MIRT models are used in this thesis.

The common MIRT models for dichotomous responses are the multidimensional two-parameter logistic (M2PL), multidimensional three-parameter logistic (M3PL) and multidimensional four-parameter logistic (M4PL) model. The probability of a correct response to item i by an examinee with ability vector $\boldsymbol{\theta}$ is mathematically modeled by the following item response functions:

$$p_i(\boldsymbol{\theta}) = P(Y = 1 | \boldsymbol{\theta}; \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(\mathbf{a}_i^T \boldsymbol{\theta} + d_i)]}, \quad (2.6)$$

M3PL

$$p_i(\boldsymbol{\theta}) = P(Y = 1 | \boldsymbol{\theta}; \mathbf{a}_i, d_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp[-(\mathbf{a}_i^T \boldsymbol{\theta} + d_i)]}, \quad (2.7)$$

M4PL

$$p_i(\boldsymbol{\theta}) = P(Y = 1 | \boldsymbol{\theta}; \mathbf{a}_i, d_i, c_i, g_i) = c_i + \frac{g_i - c_i}{1 + \exp[-(\mathbf{a}_i^T \boldsymbol{\theta} + d_i)]}. \quad (2.8)$$

In these equations, $p_i(\boldsymbol{\theta})$ is a probability of correctly response to item i by examinee with m -dimensional ability vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m) \in \Theta = \mathbb{R}^m$, $\mathbf{a}_i = (a_{i1}, \dots, a_{im}) \in \mathbb{R}^m$ is a vector of discrimination parameters for an item i , $d_i \in \mathbb{R}$ is the scale difficulty parameter of item i , $c_i \in [0, 1)$ is the guessing parameter of item i , $g_i \in (0, 1]$ is the upper/right asymptote of item i and $\mathbf{a}_i^T \boldsymbol{\theta} = \sum_{l=1}^m a_{il} \theta_l$.

For $m = 1$, $d_i = a_i b_i$.

The MIRT model for ordered categorical responses is the multidimensional graded response (MGR) model. In multidimensional context, the probability of the examinee to respond the k^{th} out of r_i categories for item i is modeled as

$$\begin{aligned}
 p_{ik}(\boldsymbol{\theta}) = P(Y = k | \boldsymbol{\theta}; \mathbf{a}_i, d_i) &= P(Y \geq k | \boldsymbol{\theta}; \mathbf{a}_i, d_i) - P(Y \geq k + 1 | \boldsymbol{\theta}; \mathbf{a}_i, d_i), \\
 P(Y \geq k | \boldsymbol{\theta}; \mathbf{a}_i, d_i) &= \frac{1}{1 + \exp[-(\mathbf{a}_i^T \boldsymbol{\theta} + d_{ik})]}, k = 1, 2, \dots, r_i - 1, \\
 P(Y \geq 0 | \boldsymbol{\theta}; \mathbf{a}_i, d_i) &= 1, \\
 P(Y \geq r_i | \boldsymbol{\theta}; \mathbf{a}_i, d_i) &= 0,
 \end{aligned} \tag{2.9}$$

where $p_{ik}(\boldsymbol{\theta})$ is a probability of correctly response to k^{th} category of item i by examinee with m -dimensional ability vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m) \in \Theta = \mathbb{R}^m$, $\mathbf{a}_i = (a_{i1}, \dots, a_{im})$ is a vector of discrimination (slope) parameter of item i and $d_i = (d_{i1}, \dots, d_{i(r_i-1)})$ is a vector of category intercepts for item i . For a comprehensive introduction of dichotomous and polytomous MIRT models, see (Reckase, 2009). The method for estimation of parameters of models mentioned in Section 2.1 and 2.2 is discussed in Chalmers (2012).

In the first paper, we have real data from an achievement test. We apply these models to this achievement test data and investigate the goodness of fit of these models by using different approaches. We also investigate how dimensionality effect the estimated parameters specially the discrimination parameter. To reach the general conclusion about the model selection and dimensionality effect, we use a case study and conduct a simulation study.

3. Optimal design theory for item calibration

There are two types of design problem in educational testing. The first is called test design in which for precise estimation of student proficiency we want to make the optimal selection of items. The second type of design is known as the sampling or calibration design problem in which for optimal estimation of the item parameters, we have to do sampling of test-takers. In educational testing, Buyske (2005) gives a review of calibration and test designs. In this thesis, we are interested in the second type of design problem.

3.1 Optimal unrestricted design

We first consider the case when we have no restrictions for availability of examinees with specific ability levels which is called optimal unrestricted design. The space of examinees' abilities is $\Theta = \mathbb{R}$. We are interested in continuous designs (see Chapter 9 in Atkinson et al. (2007)). Suppose we want to calibrate n different items and each examinee can calibrate at most one item.

We represent designs by probability-measures ξ over the design space $\chi = \Theta \times \{1, \dots, n\}$. A $(\theta, i) \in \chi$ means here that examinees with ability θ are sampled for item i . The restriction ξ_i of ξ to $\Theta \times \{i\}$ describes how abilities of examinees should be chosen for item i . The designs can be formulated for n items as

$$\begin{aligned} \xi_1 &= \left\{ \begin{array}{cccc} \theta_{11} & \theta_{12} & \dots & \theta_{1m_1} \\ w_{11} & w_{12} & \dots & w_{1m_1} \end{array} \right\} \\ \xi_2 &= \left\{ \begin{array}{cccc} \theta_{21} & \theta_{22} & \dots & \theta_{2m_2} \\ w_{21} & w_{22} & \dots & w_{2m_2} \end{array} \right\} , & \quad \begin{array}{l} 0 \leq w_{ij} \leq 1 \\ \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij} = 1 \end{array} \\ \vdots & \\ \xi_n &= \left\{ \begin{array}{cccc} \theta_{n1} & \theta_{n2} & \dots & \theta_{nm_n} \\ w_{n1} & w_{n2} & \dots & w_{nm_n} \end{array} \right\} \end{aligned}$$

where w_{ij} is a sample proportion of examinees having ability level θ_{ij} , m_i is the number of chosen distinct ability levels for item i and $\sum_{j=1}^{m_i} w_{ij}$ is a proportion

of examinees assigned to item i . To find the optimal item calibration design $\xi = (\xi_1, \xi_2, \dots, \xi_n)$, we need Fisher's information matrix for item parameters which is a block diagonal matrix defined as

$$M(\xi) = \text{diag}(M_1(\xi_1), \dots, M_n(\xi_n))$$

The models described in Chapter 3 are all generalized linear models (GLM) and the logit link is defined by

$$\eta_i(\theta) = \log \left(\frac{p_i(\theta)}{1 - p_i(\theta)} \right). \quad (3.1)$$

The standardized information matrix for the i^{th} item with item parameter β_i is expressed as

$$M_i(\xi_i) = \sum_{j=1}^{m_i} w_{ij} p_i(\theta) (1 - p_i(\theta)) \left(\frac{\partial \eta(\theta)}{\partial \beta_i} \right) \left(\frac{\partial \eta(\theta)}{\partial \beta_i} \right)^T,$$

see e.g. Atkinson et al. (2007).

In order to search an optimal design, we need to optimize some appropriate convex function Ψ of $M(\xi)$. A design ξ^* is called Ψ -optimal if $\xi^* = \arg \min_{\xi} \Psi\{M(\xi)\}$. The information matrix $M(\xi)$ contains unknown item parameters. In the optimization process we use some guess values from experts in place of the item parameters. Consequently, the design will be locally Ψ -optimal (Atkinson et al., 2007) in the sense that it is optimal for specific values of item parameters.

The directional derivative which is an important part of the General Equivalence Theorem (see below) tells how the information of a design ξ changes in a direction of another design λ :

$$F_{\Psi}(\xi, \lambda) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} [\Psi\{M((1 - \alpha)\xi + \alpha\lambda)\} - \Psi\{M(\xi)\}].$$

Let $\delta_{(\theta, i)}$ be the measure when all observations are made for ability θ and item i . We write $F_{\Psi}(\xi, \theta, i) = F_{\Psi}(\xi, \delta_{(\theta, i)})$. An optimality criterion Ψ is called differentiable if all directional derivatives can be expressed as integral over directional derivatives with respect to $\delta_{(\theta, i)}$: $F_{\Psi}(\xi, \lambda) = \int_{\mathcal{X}} F_{\Psi}(\xi, \theta, i) \lambda(d(\theta, i))$, see e.g. Whittle (1973). We assume in this thesis that criterion Ψ is differentiable. Several optimality criteria Ψ (e.g. A-, E-, L- and D-optimality) have been proposed in literature. The L-, A- and D-optimality are differentiable among the mentioned criteria. We use the D-optimality criterion in our examples which is computed as

$$\text{Minimize : } \Psi\{M(\xi)\} = -\log |M(\xi)| = -\sum_{i=1}^n \log |M_i(\xi_i)|$$

A General Equivalence Theorem was introduced by Kiefer and Wolfowitz (1960) to check whether a design is Ψ -optimal among all designs. Later Silvey (1980) alternatively formulated it by using a directional derivative. It states the equivalence of the following three conditions for the design ξ^* :

- The design ξ^* minimizes $\Psi\{M(\xi)\}$.
- The minimum over $(\theta, i) \in \chi$ of $F_\Psi(\xi^*, \theta, i) \geq 0$.
- The minimum over $(\theta, i) \in \chi$ of $F_\Psi(\xi^*, \theta, i) = 0$ and it is achieved at the support-points (θ, i) of the design ξ^* .

By plotting the directional derivative, one could check the optimality of a design.

3.2 Example

Suppose we want to calibrate items using an unrestricted design. We assume that the items follow a famous 2PL model (2.2). For this 2PL model, we present link function, derivative, information matrix, D-optimality criterion and directional derivative for D-optimality:

The link function is: $\eta_i(\theta) = a_i(\theta - b_i)$.

The derivative of the link function with respect to β_i is:

$$\frac{\partial \eta_i(\theta)}{\partial \beta_i} = \begin{bmatrix} \frac{\partial \eta_i(\theta)}{\partial a_i} \\ \frac{\partial \eta_i(\theta)}{\partial b_i} \end{bmatrix} = \begin{bmatrix} (\theta - b_i) \\ -a_i \end{bmatrix} \text{ where } \beta_i = (a_i, b_i).$$

The information matrix is:

$$M_i(\xi_i) = \sum_{j=1}^{m_i} w_{ij} p_i(\theta) (1 - p_i(\theta)) \begin{bmatrix} (\theta_{ij} - b_i)^2 & -a_i(\theta_{ij} - b_i) \\ -a_i(\theta_{ij} - b_i) & a_i^2 \end{bmatrix}.$$

The D-optimality criterion is :

$$\text{Minimize : } \Psi\{M(\xi)\} = -\log |M(\xi)| = -\sum_{i=1}^n \log |M_i(\xi_i)|$$

The directional derivative for this criterion is then given by

$$F_D(\xi, \theta, i) = 2 - p_i(\theta)(1 - p_i(\theta)) \begin{bmatrix} (\theta - b_i) & -a_i \end{bmatrix} M_i(\xi_i)^{-1} \begin{bmatrix} (\theta - b_i) & -a_i \end{bmatrix}^T.$$

The D-optimal unrestricted design for a given item following a 2PL model with item parameter $\beta_i = (a_i, b_i)$ has equal weight at two design points θ_1 and θ_2 such that the probability to answer the item at these points are $p(\theta_1) = 0.176$ and $p(\theta_2) = 0.824$. Explicitly, the two points ability design is $\theta_i = b_i \pm \frac{1.543}{a_i}$ (Abdelbasit and Plackett, 1983)

Suppose $n = 2$ and we want to calibrate Item 1 ($a_1 = 1, b_1 = 0.5$) and Item 2 ($a_2 = 1.5, b_2 = -1.2$) by using 80% of the population of examinees. The unrestricted design suggests to sample 20% at each ability levels -1.043, 2.043 (for Item 1) and -2.229, 2.043 (for Item 2). The directional derivative plot in the lower panel of Figure 3.1 confirms that the design with these points is optimal since we calibrate the items with examinees having ability levels where the directional derivatives of the items touch the reference line. The blue reference line is corresponding to the value of zero in the Generalized Equivalence Theorem for item calibration for unrestricted optimal design.

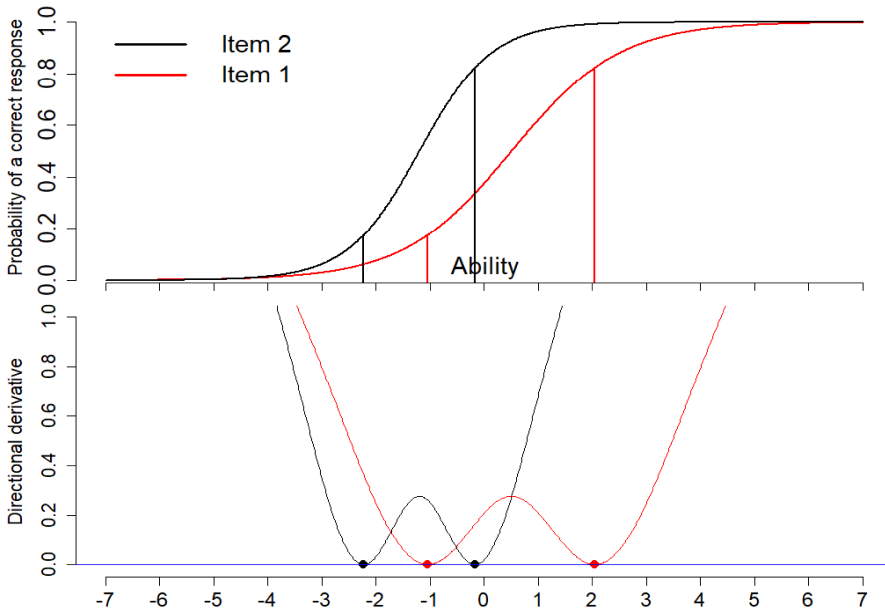


Figure 3.1: Calibration of Item 1 and 2 using an unrestricted design

The problem with the unrestricted design is to select a sample of examinees with these specific ability levels as there might not be such examinees available or we have a limited number of examinees with these ability levels. So, we have to select examinees around these abilities points. We sample instead examinees from the available distribution in an optimal way using the restricted optimal design described in Section 3.3.

3.3 Optimal restricted design

We sample examinees at some specific ability levels if we use optimal unrestricted design. Now, we are looking for intervals of ability levels to sample the examinees for optimal item calibration. For this purpose, we use an optimal restricted design and describe an Equivalence Theorem for item calibration to verify the optimality of our restricted design.

In restricted design we assume that the population of examinees (the examinees participating in the computerized achievement test) follows a continuous density g on $\Theta = \mathbb{R}^m$. For $m=1$, we assume g is a any univariate continuous density otherwise g is a any multivariate continuous density.

A sampling design is described by sub-densities $h_i \geq 0$ on Θ , $i = 0, 1, 2, \dots, n$. These sub-densities h_i together represent the whole population of examinees g available in computerized achievement test i.e.

$$\sum_{i=0}^n h_i(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \Theta, \quad (3.2)$$

where h_1, \dots, h_n represent the sub-population of examinees to be assigned to item $1, \dots, n$, respectively and h_0 describes the non-sampling distribution.

Naturally, we would use all the available population of examinees in an achievement test for calibration, but to describe the method in general term, we allow to use a sub-population of examinees. We introduce a proportion $s \in (0, 1]$ of examinees for item calibration. The value of s can be 1 if there are several items to calibrate i.e. $n \geq 2$. This restriction to a proportion s means for the non-sampled population h_0 :

$$\int_{\Theta} h_0(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 - s. \quad (3.3)$$

This non-sampled region becomes zero if we use the whole population of examinees for calibration. We define a single density h on $\chi = \Theta \times \{0, 1, \dots, n\}$ by defining $h(\boldsymbol{\theta}, i) = h_i(\boldsymbol{\theta})$. The density $h_i(\boldsymbol{\theta})$ on χ represents the density of ability level $\boldsymbol{\theta}$ of examinees to calibrate item i . The density h is a probability measure with

$$\int_{\chi} h(\boldsymbol{\theta}, i) d(\boldsymbol{\theta}, i) = 1.$$

The set of all designs h with property (3.2) and (3.3) is denoted as Ξ_s^g . The models described in Chapter 3 are all generalized linear model (GLM) and the logit link is defined by

$$\eta_i(\boldsymbol{\theta}) = \log \left(\frac{p_i(\boldsymbol{\theta})}{1 - p_i(\boldsymbol{\theta})} \right). \quad (3.4)$$

The standardized information matrix for the i^{th} item with continuous density $h_i(\boldsymbol{\theta})$ and item parameter $\boldsymbol{\beta}_i$ is expressed as

$$M_i(h_i) = \int_{\Theta} P_i(\boldsymbol{\theta})(1 - p_i(\boldsymbol{\theta})) \left(\frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_i} \right) \left(\frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_i} \right)^T h_i(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (3.5)$$

To search optimal restricted designs, we need the total information matrix $M(h)$ which is a block-diagonal matrix for a model of n items with item parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n)$ and continuous density $h = (h_1, \dots, h_n)$ is defined as

$$M(h) = \text{diag}(M_1(h_1), \dots, M_n(h_n)).$$

In order to find an optimal restricted design in the set of restricted designs Ξ_s^g , we need to optimize an appropriate convex function Ψ of $M(h)$. A design h^* is Ψ -optimal under restriction (3.2) and (3.3) if $h^* = \arg \min_h \Psi\{M(h)\}$. The models defined in Chapter 2 are nonlinear. In the nonlinear case, the information matrix $M(h)$ depends on the model parameters $\boldsymbol{\beta}$. These item parameters are unknown. To find optimal design, we need some prior knowledge or appropriate guess about item parameters. So, the optimal design h^* is a locally optimal restricted design.

The optimality criteria are some appropriate functions of $M(h)$. The D-optimality is a famous and the most often applied criterion to online items calibration literature (Chang and Lu, 2010; Jones and Jin, 1994; Zhu, 2006). The D-optimality criterion is computed as

$$\text{Minimize : } \Psi\{M(h)\} = \log |M^{-1}(h)| = -\log |M(h)| = -\left(\sum_{i=1}^n \log(|M_i(h_i)|) \right). \quad (3.6)$$

The directional derivatives quantify the change in criterion function Ψ of $M(h)$ if a small amount of observations in $\boldsymbol{\theta}$ are added to the i^{th} item. The directional derivative of Ψ at $M(h)$ in the direction of the one-point measure $\delta_{(\boldsymbol{\theta}, i)}$ in $(\boldsymbol{\theta}, i)$ is defined as

$$F_{\Psi}(h, \boldsymbol{\theta}, i) = F_{\Psi}(h, \delta_{(\boldsymbol{\theta}, i)}) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} [\Psi\{M((1 - \alpha)h + \alpha \delta_{(\boldsymbol{\theta}, i)})\} - \Psi\{M(h)\}].$$

The directional derivative and minimization function is similar as in Section 3.1. The only difference is that now we have density h instead of design measure ξ . The directional derivative is an important part of the General Equivalence Theorem, which we use for characterization an optimal design. We

derive a new equivalence theorem for the case of calibration of multiple items. This Equivalence Theorem uses the directional derivative $F_{\Psi}(h, \boldsymbol{\theta}, i)$ for the design h when a small amount of observations in $\boldsymbol{\theta}$ are added to the i^{th} item. We define the minimum \tilde{L} over the directional derivatives for the given design h^* as

$$\tilde{L}(h^*, \boldsymbol{\theta}) = \min_{i=1, \dots, n} F_{\Psi}(h^*, \boldsymbol{\theta}, i). \quad (3.7)$$

We further define for a given sampling proportion s

$$c^* = \arg \max_c \left\{ \int_{\Theta} \mathbf{1}_{\tilde{L}(h^*, \boldsymbol{\theta}) \leq c} g(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq s \right\}, \quad (3.8)$$

where $\mathbf{1}_A$ is the indicator function being 1 on a set A and 0 otherwise. Let L be the at c^* and truncated function \tilde{L} , $L(h^*, \boldsymbol{\theta}) = \min\{\tilde{L}(h^*, \boldsymbol{\theta}), c^*\}$. When formally defining $F_{\Psi}(h^*, \boldsymbol{\theta}, 0) = c^*$, we can write

$$L(h^*, \boldsymbol{\theta}) = \min_{i=0, 1, \dots, n} F_{\Psi}(h^*, \boldsymbol{\theta}, i). \quad (3.9)$$

Theorem 1 (Equivalence Theorem for Item Calibration) *Let $h^* \in \Xi_s^g$ be a design and c^* and L be defined according to (3.8) and (3.9) and let Ψ be differentiable. Then: h^* is Ψ -optimal in Ξ_s^g if and only if*

$$F_{\Psi}(h^*, \boldsymbol{\theta}, i) = L(h^*, \boldsymbol{\theta}) \text{ for } h^* \text{-almost all } (\boldsymbol{\theta}, i) \in \tilde{\chi}. \quad (3.10)$$

Note: In Paper II we have $\Theta = \mathbb{R}$ but as discussed in Paper IV, the theorem holds for $\Theta = \mathbb{R}^m$ and the proof is in a similar manner.

In Paper II, we develop an algorithm which directly optimizes the criterion function Ψ in order to find the optimal restricted design. We use this equivalence theorem for item calibration to check that the given design is optimal by computing and plotting the directional derivative of each item.

With the unrestricted design (Figure 3.1), it is not possible to take 20% of the students at each unrestricted design point for item calibration due to unavailability of students with such specific abilities. So, Figure 3.2 shows the restricted design for calibration of item using 80% population of examinees. According to the equivalence theorem for item calibration, it is an optimal restricted design.

The algorithm used in Paper II becomes complicated with the increase of the number of items to be calibrated. In Paper III, we develop an exchange algorithm for restricted optimization when a finite population is available. An exchange algorithm has been suggested of Fedorov (1989) for sampling a sub-population. Our algorithm discretizes the ability space into small intervals and iteratively exchanges these small intervals. It is built on the equivalence theorem. In this paper, we also develop some asymptotic theorems which help us

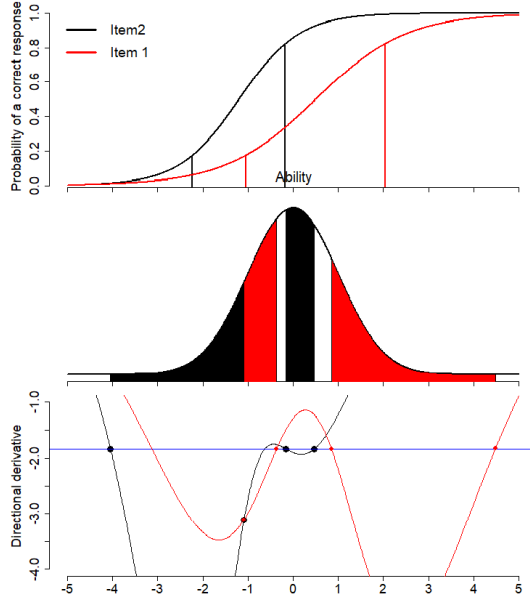


Figure 3.2: Calibration of Item 1 and 2 by 80% of the examinees. Upper panel: black and red lines represent the two-parameter logistic model curves. The vertical lines indicate the ability levels of the locally D-optimal unrestricted design; Middle panel: two shaded red and black parts of normal distribution represent the intervals for ability levels of the locally D-optimal restricted design for these items; Lower panel: Black and red line curves represent the directional derivatives while four black and red dots marked on it depict the lower and upper interval limits of the locally D-optimal restricted design of these items. The blue line is a reference line and where we sample when the directional derivative is below in this line.

to choose an item at high ability level ($\theta \rightarrow \infty$) and low ability level ($\theta \rightarrow -\infty$) to sample the examinees. These asymptotic theorems help to speed up the algorithm and search the minimum number of intervals.

Theorem 2 *Suppose the calibration items follow a 2PL model. For the locally D-optimal restricted design, the following holds: The examinees at low and high ability levels are calibrated with an item which has the smallest discrimination parameter. For more details and the proof see Theorem 2 in Paper III.*

Theorem 3 *Suppose the calibration items follow a 3PL model. For the locally D-optimal restricted design, the following holds: The examinees at high ability levels are calibrated with an item which has the smallest discrimination*

parameter. For more details and the proof see Theorem 3 in Paper III.

Theorem 4 *Suppose some calibration items follow a 2PL model and some follow a 3PL model. For the locally D-optimal restricted design, The following holds: the examinees at high ability levels are calibrated with an item which follows a 2PL model and has the smallest discrimination parameter among the 2PL items. For more details and the proof see Theorem 4 in Paper III.*

In Paper IV, we generalized the algorithm, developed in Paper II for simultaneous calibration of items in a multidimensional setting using Theorem 1. We also develop an asymptotic theorem which helps us to choose an item at extreme ability levels to sample the examinees.

Theorem 5 *Suppose we want to calibrate n items following a M2PL model with examinees having m ability dimensions $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m) \in \Theta = \mathbb{R}^m$ and we consider D-optimality. Let $a_{i1}, a_{i2}, \dots, a_{im}$ be the discrimination parameters of item $i = 1, 2, \dots, n$. The optimal design samples examinees with extreme ability level ($\|\boldsymbol{\theta}\|$ large) for the item i which has the smallest value of $|\sum_{l=1}^m a_{il}\theta_l|$. For more details and the proof see Theorem 2 in Paper IV.*

We can use this asymptotic theorem result in order to speed up the algorithm. One can choose the starting design based on this theorem.

3.4 Efficiency comparison

We use relative efficiency (RE) to compare the efficiency of one design compared to another design. The relative D-efficiency (RE_D) of any design h_r compared to the D-optimal design h^* is defined as

$$RE_D = \left[\frac{|M(h_r)|}{|M(h^*)|} \right]^{\frac{1}{P}} = \left[\frac{\prod_{i=1}^n |M_i(h_{ri})|}{\prod_{i=1}^n |M_i(h_i^*)|} \right]^{\frac{1}{P}},$$

where P is the number of parameters see Berger and Wong (2009). A relative D-efficiency less than 1 indicates that design h^* is more efficient than h_r in terms of D-optimality. In terms of sample size, h_r would take about $(RE_D^{-1} - 1) * 100$ percent more examinees compared to design h^* to have the same efficiency as the design h^* .

In Paper II, we compare the relative efficiency of a random and a symmetric design compared to the D-optimal restricted design. The random design assigns items randomly irrespectively of examinees' abilities and each examinee

has an equal probability to calibrate a specific item. The symmetrical design considers the intervals of the ability levels of examinees symmetrically around the unrestricted design points. For the example of Section 3.2, the random design requires 24.43% more examinees to be as efficient as the locally D-optimal restricted design when we want to calibrate Item 1 and 2 with 80% of the available examinees (see Table 3 in Paper II). In Paper III and IV, we compare the relative efficiency of the random design compared to the D-optimal restricted design.

4. Overview of the papers and future research

Paper I: Discrimination with unidimensional and multidimensional item response theory models for educational data

In Paper I, we analyze achievement test data from higher education. We apply different IRT models to this data and use different approaches in order to search an appropriate model that characterizes the items appropriately. We investigate whether dimensionality influences the estimates of the item-parameters especially the discrimination parameter. By using the case and a simulation study, we identify that a multidimensional model better discriminates among the students.

Paper II: Optimal item calibration for computerized achievement tests

In Paper II, we describe a new method for the efficient selection of examinees' abilities for optimal item calibration in computerized achievement tests. To sample examinees with some specific abilities efficiently, we use optimal design theory and assume that the probability of a correct response follows an item response model. Locally D-optimal unrestricted designs give us a few points to sample the examinees for item calibration, but in practice it is hard to sample examinees from the population with these specific ability levels. We use the approach of locally D-optimal restricted designs which gives us some intervals of ability levels to sample the examinees from the population.

We develop an algorithm for selection of optimal intervals. We prove an equivalence theorem needed to verify the optimality of a design. By using our algorithm, many scenarios are presented for calibration of single items and simultaneous calibration of several items by assuming a 2PL model. We assume standard normal distribution for examinees' ability, but one could assume any distribution for examinees' ability here. We can use any proportion of the population of examinees in achievement test for calibration of item. Moreover, the algorithm is flexible to handle any other convex and differentiable

optimality criterion (e.g. I- and A-optimality). We also discuss how to scale up this method for a large bank of new items.

Paper III: An exchange algorithm for optimal calibration of items in computerized achievement tests

The algorithm described in Paper II optimizes the criterion directly. This algorithm becomes complicated when the number of items for calibration increases. In this paper, we have developed a new exchange algorithm for calibration of several competing items based on equivalence theorem proved in Paper II. The algorithm uses all available population of examinees in computerized achievement test for calibration of items. The algorithm is flexible to handle any other convex and differentiable optimality criterion (e.g. I- and A-optimality). It works without any problem for very general IRT models including the 1PL-, 2PL- and 3PL-model and in the situation when some items follow a 2PL-model and some a 3PL-model. We have also proved some asymptotic results to calibrate the item at extreme ability level for D-optimality when items follow different IRT models. We assume here standard normal distribution for examinees' abilities, but in principle one could assume any distribution for examinees' abilities. These asymptotic results help to construct the starting design for the algorithm, increase the speed of the algorithm and compute the optimal design with a small number of intervals in complicated situations.

Paper IV: Optimal calibration of items for multidimensional achievement tests

In Paper IV, we generalize the exchange algorithm described in Paper III for calibration of items assuming examinees in achievement test have multidimensional ability to answer the questions. The equivalence theorem proved in Paper II is also valid for arbitrary dimension m of the ability space.

We prove an asymptotic theorem using the M2PL model for an arbitrary m dimensional space of ability. This asymptotic theorem gives a solution about which item should be calibrated by examinees with extreme ability. This asymptotic theorem helps to construct a starting design and to speed up the algorithms.

In this paper for simplicity, we work with M2PL model but one could use this algorithm for other MIRT model, e.g. M3PL, M4PL and multidimensional graded response model. This algorithm is flexible to handle any other convex and differentiable optimality criterion (e.g. I- and A-optimality). The idea for

computing designs with this algorithm is valid for m dimensions. However, it would be challenging to use the algorithm with large m .

Future research

In this thesis, we use guess values of items parameter for item calibration. Consequently, an optimal item calibration design is locally optimal. Alternatively, one may use minimax or Bayesian method approach (see Atkinson et al. (2007), Chapter 17 and 18) in the future, which use prior information about the range and distribution of item parameters, respectively.

We use here estimated proficiencies of examinees for optimal calibration of items which is a latent variable and cannot be directly observed. If these estimated abilities are far from the real ones, we obtain a bad design. For future research, one may incorporate the measurement error in an ability variable in order to get an optimal restricted design.

We develop methods and algorithms for item calibration using optimality criteria which are differentiable. One may consider other optimality criteria (e.g. E-optimality) for further research.

It would also be interesting in the future to develop methods and implementation of algorithms for the testlets. For testlets, we are interested to search an optimal restricted design for certain groups of items. Each group contains items belonging to the same context or topic.

5. Sammanfattning

Kunskapsprov används för att mäta studenters färdigheter inom specifika kunskapsområden. Datorbaserade kunskapsprov (t.ex. GRE och SAT) baseras vanligen på frågor ur en uppgiftsbank för att mäta studenternas färdigheter. En uppgiftsbank är en stor samling av uppgifter med kända egenskaper (t.ex. svårighetsnivå). Uppgiftsbankerna uppdateras och revideras kontinuerligt med nya uppgifter som ersätter uppgifter som blivit förlegade, överexponerade eller bristfälliga med tiden. Denna avhandling ägnas åt att uppdatera och upprätthålla en uppgiftsbank med högkvalitativa uppgifter samt förbättrade skattningar av uppgiftsparametrar (utprovning av uppgifter).

Avhandlingen består av fyra artiklar. En artikel undersöker vilken inverkan dimensionaliteten av förmåga har på skattade parametrar medan övriga tre behandlar utprovning av uppgifter.

I den första artikeln undersöker vi hur dimensionaliteten av förmåga påverkar skattningarna av uppgiftsparametrarna. I en fall- och simuleringsstudie fann vi att en flerdimensionell modell bättre klarar av att särskilja studenter.

Den andra artikeln beskriver en metod för optimal uppgiftskalibrering genom effektiv selektion av provtagare baserat på deras kunskapsnivå. Vi utvecklar en algoritm som identifierar intervall för studentens nivå av förmåga för optimal utprovning av uppgifterna. Vi utvecklar också en ekvivalenssats för utprovning som kan användas för att verifiera det optimala designet.

Algoritmen som utvecklas i Artikel II blir mer komplicerad med ökat antal utprovningssuppgifter. Därför utvecklar vi i Artikel III en ny utbytesalgoritm baserat på ekvivalenssatsen som utvecklades i Artikel II.

Slutligen generaliserar den fjärde artikeln utbytesalgoritmen beskriven i Artikel III till antagandet att stunderna har flerdimensionella förmågor för att lösa uppgifterna.

References

- Abdelbasit, K. M. and Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association*, 78(381):90–98. 10
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573. 5
- Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press, Oxford. 7, 8, 19
- Berger, M. P. F. and Wong, W. K. (2009). *An introduction to optimal designs for social and biomedical research*, volume 83. John Wiley & Sons. 15
- 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. 5
- Buyske, S. (2005). *Optimal Designs in Educational Testing*, in *Applied Optimal Designs* (eds M. P. F. Berger and W. K. Wong). John Wiley & Sons, Ltd, Chichester, UK. 4, 7
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6):1–29. 6
- Chang, Y.-C. I. and Lu, H.-Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika*, 75(1):140–157. 12
- Drasgow, F. and Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7(2):189–199. 5
- Fedorov, V. V. (1989). Optimal design with bounded density: Optimization algorithms of the exchange type. *Journal of Statistical Planning and Inference*, 22:1–13. 13
- Jones, D. H. and Jin, Z. (1994). Optimal sequential designs for on-line item estimation. *Psychometrika*, 59(1):59–75. 12
- Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12(5):363–365. 9

- Kirisci, L., Hsu, T.-c., and Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, 25(2):146–162. 5
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174. 5
- Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, 1992(1):i–30. 5
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, 4(3):207–230. 5
- Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer. 6
- Reckase, M. D. and McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 15(4):361–373. 5
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4):100. 4
- Silvey, S. D. (1980). *Optimal Design*. Monographs on Applied Probability and Statistics 1. Chapman and Hall, London, 1st edition. 9
- van der Linden, W. J. (2016). *Handbook of item response theory, volume one: models*. CRC Press. 3, 5
- van der Linden, W. J. and Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media. 5
- Way, W. D., Ansley, T. N., and Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, 12(3):239–252. 5
- Whittle, P. (1973). Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(1):123–130. 8
- Yang, S. (2007). *A comparison of unidimensional and multidimensional rasch models using parameter estimates and fit indices when assumption of unidimensionality is violated*. PhD thesis, The Ohio State University. 5
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145. 5

- Zheng, Y. and Chang, H. H. (2017). A comparison of five methods for pretest item selection in online calibration. *International Journal of Quantitative Research in Education*, 4(1-2):133–158. 1
- Zhu, R. (2006). *Implementation of optimal design for item calibration in computerized adaptive testing (CAT)*. University of Illinois at Urbana-Champaign. 12

