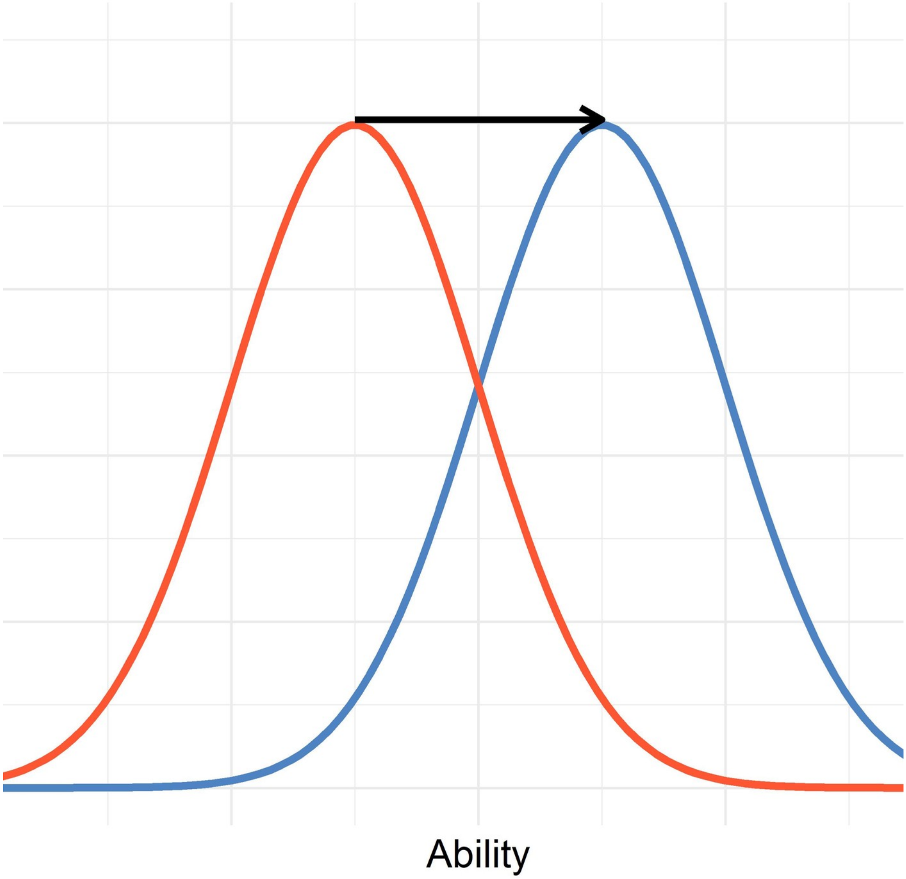
Stockholm University

Test Design for Mean Ability Growth and Optimal Item Calibration for Achievement Tests

Jonas Bjermo



Doctoral Thesis in Statistics at Stockholm University, Sweden 2021

Test Design for Mean Ability Growth and Optimal Item Calibration for Achievement Tests

Jonas Bjermo

Academic dissertation for the Degree of Doctor of Philosophy in Statistics at Stockholm University to be publicly defended on Friday 10 December 2021 at 10.00 in hörsal 4, hus 2, Albanovägen 12.

Abstract

In this thesis, we examine two topics in the area of educational measurement. The first topic studies how to best design two achievement tests with common items such that a population mean-ability growth is measured as precisely as possible. The second examines how to calibrate newly developed test items optimally. These topics are two optimal design problems in achievement testing. Paper I consist of a simulation study where different item difficulty allocations are compared regarding the precision of mean ability growth when controlling for estimation method and item difficulty span. We take a more theoretical approach on how to allocate the item difficulties in Paper II. We use particle swarm optimization on a multi-objective weighted sum to determine an exact design of the two tests with common items. The outcome relies on asymptotic results of the test information function. The general conclusion of both papers is that we should allocate the common items in the middle of the difficulty span, with the two separate test items on different sides. When we decrease the difference in mean ability between the groups, the ranges of the common and test items coincide more.

In the second part, we examine how to apply an existing optimal calibration method and algorithm using data from the Swedish Scholastic Aptitude Test (SweSAT). We further develop it to consider uncertainty in the examinees' ability estimates. Paper III compares the optimal calibration method with random allocation of items to examinees in a simulation study using different measures. In most cases, the optimal design method estimates the calibration items more efficiently. Also, we can identify for what kind of items the method works worse.

The method applied in Paper III assumes that the estimated abilities are the true ones. In Paper IV, we further develop the method to handle uncertainty in the ability estimates which are based on an operational test. We examine the asymptotic result and compare it to the case of known abilities. The optimal design using estimates approaches the optimal design assuming true abilities for increasing information from the operational test.

Keywords: *test design, item response theory, optimal experimental design, SweSAT, item calibration, vertical scaling, ability growth, computerized adaptive tests.*

Stockholm 2021
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-197928

ISBN 978-91-7911-674-3
ISBN 978-91-7911-675-0

Stockholm University

Department of Statistics

Stockholm University, 106 91 Stockholm

TEST DESIGN FOR MEAN ABILITY GROWTH AND OPTIMAL ITEM CALIBRATION FOR ACHIEVEMENT TESTS

Jonas Bjermo

Stockholm University

Test Design for Mean Ability Growth and Optimal Item Calibration for Achievement Tests

Jonas Bjermo

©Jonas Bjermo, Stockholm University 2021

ISBN print 978-91-7911-674-3
ISBN PDF 978-91-7911-675-0

Printed in Sweden by Universitetsservice US-AB, Stockholm 2021

To my parents.

Abstract

In this thesis, we examine two topics in the area of educational measurement. The first topic studies how to best design two achievement tests with common items such that a population mean-ability growth is measured as precisely as possible. The second examines how to calibrate newly developed test items optimally. These topics are two optimal design problems in achievement testing.

Paper I consist of a simulation study where different item difficulty allocations are compared regarding the precision of mean ability growth when controlling for estimation method and item difficulty span. We take a more theoretical approach on how to allocate the item difficulties in Paper II. We use particle swarm optimization on a multiobjective weighted sum to determine an exact design of the two tests with common items. The outcome relies on asymptotic results of the test information function. The general conclusion of both papers is that we should allocate the common items in the middle of the difficulty span, with the two separate test items on different sides. When we decrease the difference in mean ability between the groups, the ranges of the common and test items coincide more.

In the second part, we examine how to apply an existing optimal calibration method and algorithm using data from the Swedish Scholastic Aptitude Test (SweSAT). We further develop it to consider uncertainty in the examinees' ability estimates. Paper III compares the optimal calibration method with random allocation of items to examinees in a simulation study using different measures. In most cases, the optimal design method estimates the calibration items more efficiently. Also, we can identify for what kind of items the method works worse.

The method applied in Paper III assumes that the estimated abilities are the true ones. In Paper IV, we further develop the method to handle uncertainty in the ability estimates which are based on an operational test. We examine the asymptotic result and compare it to the case of known abilities. The optimal design using estimates approaches the optimal design assuming true abilities for increasing information from the operational test.

Keywords: ability growth, computerized adaptive tests, item calibration, item response theory, optimal experimental design, SweSAT, test design.

List of Papers

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

PAPER I: Jonas Bjermo, Frank Miller, Efficient Estimation of Mean Ability Growth Using Vertical Scaling, *Applied Measurement in Education*, **34(3)**: 163-178 (2021).
DOI: 10.1080/08957347.2021.1933981

PAPER II: Jonas Bjermo, Optimal Test Design for Estimation of Mean Ability Growth.
(Manuscript)

PAPER III: Jonas Bjermo, Ellinor Fackle-Fornius, Frank Miller, Optimal Item Calibration in the Context of the Swedish Scholastic Aptitude Test.
(Submitted)

PAPER IV: Jonas Bjermo, Ellinor Fackle-Fornius, Frank Miller, Optimizing Calibration Designs with Uncertainty in Abilities.
(Manuscript)

Reprints were made with permission from the publishers.

Acknowledgements

As a child, I was always curious. I spent a lot of time observing different kinds of natural phenomena, such as the beautiful starry sky, the behavior of animals, and other things. I also liked dismantling technical devices to understand how they worked, not always managing to assemble them again. I was probably considered to be a little bit odd to my surroundings. But my parents and brother let me be who I am, for which I am eternally grateful. This has led me to where I am in my career today.

In writing this thesis, I am very grateful for the help and support of my main supervisor Frank Miller for providing ideas and always answering my questions. I am also very grateful to my supervisor Ellinor Fackle-Fornius, who has also been very supportive. I have received a lot of guidance and encouragement from both of you. Your help has not only made this thesis possible but also made me a more proficient statistician.

I would also like to thank Oscar that has been around since our time at the National Agency for Education. You always gave me mathematical puzzles that distracted me from work. We have also relieved some stress by lifting heavy weights with Oskar, often while discussing interesting statistical issues.

Renata and I started the Ph.D. journey simultaneously, and we have been helping each other through the process. I am also grateful to all who have been Ph.D. students at the department at the same time. Thank you for your support. The rest of the staff at the department have been contributed to the creative and friendly atmosphere, which has facilitated the writing of this thesis. I have really enjoyed my time at the department.

I would also like to mention my former colleagues at the National Agency of Education which have contributed to a more general understanding of the educational system and large-scale assessments. Some of the ideas of this dissertation started when working with you.

Finally, I would like to give my deepest love to my wonderful wife, Jenny. Writing this thesis has been a lot easier with you on my side. You make my life more enjoyable.

Contents

1. Introduction

1.1 Background

Achievement tests are used all over the world in many different contexts. It is often of interest to make comparisons of achievements in school subjects between students, schools, and countries. The Trends In International Mathematics And Science Study (TIMSS) (International Study Center, Boston College, 2017) and the Programme for International Student Assessment (PISA) (OECD, 2019) compare student abilities on a country level. The Swedish national tests (Swedish National Agency for Education, 2019) are a support for the teacher to give the students a fair assessment. Individual test results are utilized to ensure adequate qualifications. Certification for university studies, such as the Swedish Scholastic Aptitude Test (SweSAT) (Umeå University, 2021) or to be certified for a driver's license (Trafikteori Online, 2021) are two examples. No matter if the purpose of the test is to assure competence at a certain level or to compare groups, the reliability of the results is of great importance.

Educational measurement is the research area that approaches questions related to testing abilities of different kinds. The assigning of numerals to traits is a frequent definition of the field. It is often of interest to estimate student abilities, such as reading skills or mathematics proficiency.

Traditionally, achievement tests have been paper-and-pencil tests with a fixed test length, being the same for all examinees. Nowadays, computers occur more often in test situations, which enable more advanced tests such as computerized adaptive tests (CAT). These tests are adapted such that the questions assigned to examinees during the test depend on their abilities.

By reliability of results, we mean good precision of the examinees' estimated abilities and the estimated parameters that characterize the test questions (also called items). It is also of interest to minimize the cost for a given precision. The cost here means the number of examinees or items selected. Depending on if the interest lays in the precision of the abilities or item parameters we distinguish between test design and calibration design. The former concerns optimally selecting items for efficient estimation of abilities and the latter concerns optimally selecting examinees for efficient estimation of the item parameters. Efficiently means minimizing the variance of the estimated parameters. The branch in statistics dealing with such situations is called optimal design. We consider both test designs and calibration designs in this thesis.

To examine the test design and item calibration in the context of this thesis, we use several methods and concepts. Some of them are Item Response Theory (IRT), Optimal Design, Scaling, and the concept of Latent Variables. The upcoming chapters will present them as preparation. The aim is to give enough understanding of the concepts used in the research papers.

1.2 The objective the of thesis

This thesis considers two kinds of optimal design in achievement testing. We apply both test design and item calibration. One part of the thesis examines and develops optimal test designs to maximize the precision of estimated mean ability growth between grades in school. We investigate how the allocation of the item difficulty parameters (test design) affects the precision of the estimated difference of means between two ability distributions. The difference is assumed to be the mean growth in ability or knowledge of a population between school grades. The optimal test design for mean ability growth is considered in Paper I and II. Paper I investigates the impact of different test designs on the precision of the estimated mean ability growth when controlling for the estimation method. We take a more theoretical approach on how to allocate the item difficulties in Paper II.

The second part considers item calibration. Here we apply and further develop a method by Ul Hassan and Miller (2019) under various realistic settings. Paper III compares an optimal calibration method with a random allocation of items to examinees in a simulation study using different measures. We use the exchange algorithm by Ul Hassan and Miller (2021) on a simulation study and apply it to Swedish Scholastic Aptitude Test (SweSAT) data (Umeå University (2021)). The method applied in Paper III assumes that the estimated abilities are the true ones. In Paper IV, we hence develop the item calibration method by Ul Hassan and Miller (2019) to handle uncertainty in the ability estimates.

1.3 Outline

The ability of an examinee, being a non-observable variable, is present throughout the whole thesis. Therefore, we start with a short introduction to latent variables in the next chapter.

The Item Response Theory (IRT) model is a statistical model connecting the items of a test and the latent abilities of the examinees. We will use the IRT model in all papers, and Chapter 3 will introduce the reader to the IRT model.

The main topic of the papers is optimal design in the context of test design and item calibration. We introduce general optimal design theory in Chapter 4 and then present how to apply it for designing achievement tests and calibrating items.

The main objective of the papers considering test design in this thesis is estimating mean ability growth between school grades. We, therefore, need to consider vertical scaling that guarantees that abilities will be on the same measurement scale. We will introduce the reader to equating and vertical scaling in Chapter 5, used for measuring ability growth in the first two papers. The thesis ends with a summary of the included papers.

2. Latent variable models

In educational and psychological measurement, a person's trait is often of interest. In a psychological context, it could be stress or motivation. In education, it could be scholastic ability. A personal trait can not be measured directly, unlike the weight or height of a person. It is therefore often referred to as a latent trait or a latent variable. A latent variable could also mean a physically existent variable that is difficult to measure. In this thesis, the latent variable represents a hypothetical construct that is difficult to measure directly, such as the ability in mathematics or reading skills.

There exist several definitions of latent variables in the literature (e.g. Skrondal and Rabe-Hesketh (2004) and Fox (2010) p. 5). It is common to define a latent variable as a random variable that cannot be directly measured but is related to observed data. An appropriate mathematical model serves as a link between the latent variable and the data. We define a latent variable as follows.

Definition 1 *A latent variable is a random variable that is not observed directly but manifests itself as an observable variable through a mathematical model. The observed variables are called a manifest variables.*

Figure 2.1 visualizes the concept of a latent variable model. A latent variable measured by three observed manifest variables with error terms. The latent variable could for example be stress, measured by three manifest variables in form of questions on a questionnaire.

Example 1 *Assume that a mathematics test has questions scored as one or zero (correct or not). A mathematical model will be selected to serve as a link between the test score and the ability of the student. The response to the question serves as an indicator of the student's ability in mathematics. A proficient student will have a higher probability of answering correctly. Often, a logistic mathematical model will serve as the link between the test score and ability.*

A latent variable can be defined as continuous or categorical. The latent variable is assumed to continuous in this thesis. In educational measurement,

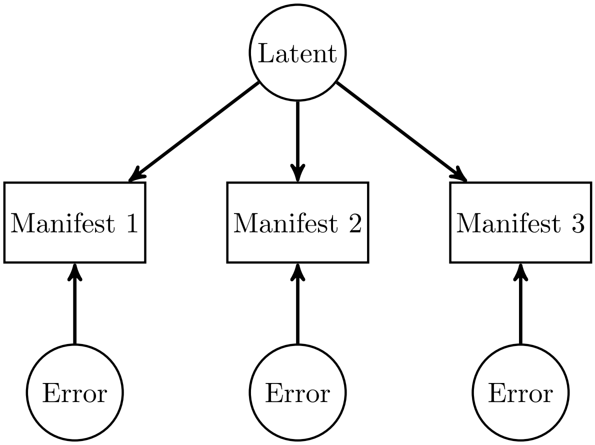


Figure 2.1: A latent variable model with three manifest variables measured with error terms.

the latent variable represents cognitive abilities such as reading or arithmetic skills. Therefore, we will refer to the latent variable as the ability of a student or examinee. We will denote it as θ in this thesis.

The ability of an examinee will manifest itself through the test score. The questions in the test are usually called items. In the next section, we will define the theoretical relationship between the score of the items and the examinee's ability. The mathematical link between the ability and the observed data could be of various forms. In this thesis, we will use the theory of Item Response Models.

3. Item response theory

There are several different statistical models for connecting the item characteristics and the examinees' ability. Item response theory (IRT) is a class of latent variable models used for measuring a latent trait measured indirectly via manifest variables. The theory describes the relationship between an examinee's performance and latent ability. The ability θ is considered to be proficiencies in reading, mathematics, or similar. The name, IRT, comes from the fact that the items are in focus compared to classical test theory that focuses on the whole test. In classical test theory, the observed test score consists of a true score and an error score (van der Linden and Hambleton, 1997). We will consider the ability of an examinee to be unidimensional. Otherwise, the item response model is said to be multidimensional (Reckase, 2009).

Items can be of different kinds. In the context of achievement tests, it is common to use dichotomously scored multiple-choice items or constructed-response items. A constructed-response item requires the student to generate a response as opposed to a multiple-choice item where a response is selected among several given alternatives. A multiple-choice item is scored either 0 or 1, depending on if the response is correct or not. A constructed-response item is scored 0 for an incorrect response and up to a maximum of $m \in \mathbb{Z}^+$. We will assume $m = 1$, which means that both multiple- and constructed response items are dichotomously scored. If the item scores take values $m > 1$, we need to use polytomous IRT models. See e.g. Nering and Ostini (2010) for a description of polytomous IRT models.

An important concept in IRT is the item characteristic curve (ICC). It is the plot of the probability of a correct answer against the ability of an examinee. Figure 3.1 shows an example of an item characteristic curve for a dichotomous scored item. The graph displays that the higher the ability, the higher probability of a correct answer. The item characteristic curve is the basic building block of item response theory; all the other constructs of the theory depend upon this curve (Baker and Kim, 2004).

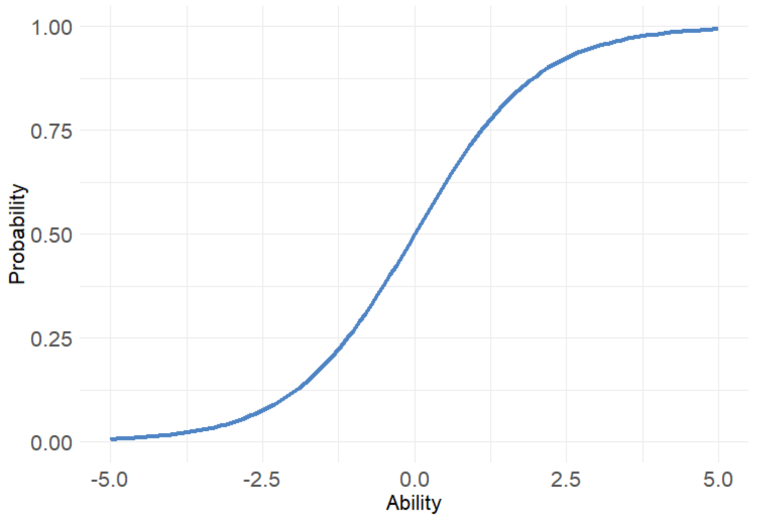


Figure 3.1: The ICC of a dichotomous scored item, which is the probability of a correct answer as a function of ability.

3.1 Model

We will now describe the relationship between the responses to dichotomous scored items and examinee ability in more detail. IRT models the probability of a correct response to an item as a function of examinee ability. When applying IRT on achievement tests in this thesis, we consider a test of n items and N examinees.

Definition 2 *Firstly, let X_{ij} be the score of examinee $j \in \{1,\ldots,N\}$ to item $i \in \{1,\ldots n\}$, scored either 0 or 1. Then the item response function is $P_i(\theta_j) := Pr(X_{ij} = 1|\theta_j)$, where θ_j is the ability of examinee j.*

Remark 1 *Since $P_i(\theta)$ is the probability of a correct response, $1 - P_i(\theta)$ is the probability of an incorrect response.*

The item response function $P_i(\theta)$ is also called the *item characteristic curve* (ICC) and can be plotted over the distribution of θ as we have already seen in Figure 3.1. Usually, we make three assumptions in IRT (Johnson, 2007).

1. The latent trait, or the ability of an examinee is one-dimensional. This is referred to as *Unidimensionality* (U).

2. The responses of the items are independent, given the examinees' ability θ_j. This is referred to as *Local Independence* (LI).

3. The ICC for all items is a non-decreasing function of examinee ability θ_j. This is referred to as *Monotonicity* (M).

The unidimensional assumption assures that the items in a test measure a single construct, which might be hard to achieve to a full extent. Often, it is common to use factor analysis to test the assumption. Local independence means that if the dependence on ability were removed in the response function, there is no longer any relationship between the test items. With monotonicity, the probability of a correct score of an item can not decrease when increasing the ability. The probability of a correct score will be higher for more proficient examinees. These assumptions make the ICC be the basic building block of IRT. Figure 3.2 shows three different ICC with different characteristics.

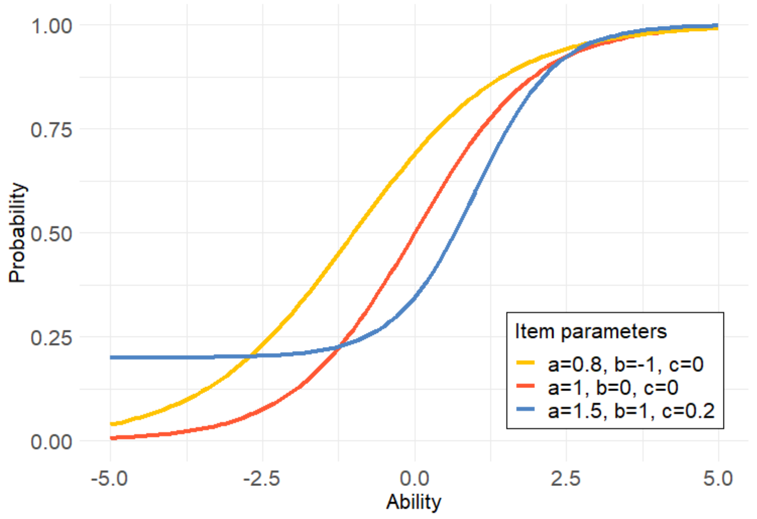


Figure 3.2: Item Characteristic Curves of three items with different item parameters.

The ICC can be modeled either as a logistic or probit curve (see e.g. Baker and Kim (2004)), where the logistic model is more common and will be considered here. Except for the ability, several item parameters will also affect the probability of a correct item response. The item parameters in a logistic model are the discrimination parameter (a), the difficulty parameter (b), and the guessing parameter (c). They are also called the slope, location, and lower asymptotic parameter. The next definition clarifies the logistic model.

Definition 3 *The logistic item characteristic curve or probability function for item i and ability θ_j is defined as* $p_{ij}(\theta_j|a_i,b_i,c_i) = c_i + \frac{1-c_i}{1+e^{-a_i(\theta_j-b_i)}}$, *where* $a_i \in (0,\infty)$, $b_i \in (-\infty,\infty)$ *and* $c_i \in [0,1]$.

Remark 2 *In practice the discrimination parameter a typically ranges between* 0.5 *and* 2*, the difficulty parameter b typically ranges between* -3 *and* 3*, and the guessing parameter c ranges between* 0 *and* 0.5*.*

The model in definition 3 is called the three parameter logistic (3PL) model. If we let $c_i = 0$, the model reduces to the two parameter logistic (2PL) model, and if we also let $a_i = 1$ it reduces to the one parameter logistic (1PL) model. Therefore, $\lim_{\theta\to\infty} P_i(\theta) = 1$ for all three models, $\lim_{\theta\to-\infty} P_i(\theta) = 0$ for the 1PL and 2PL model, and $\lim_{\theta\to-\infty} P_i(\theta) = c$ for the 3PL model.

The 3PL model is often used for multiple-choice items when there is a possibility to guess the correct answer. For dichotomous scored constructed-response items, the 2PL model is more suitable since it is very unlikely to give a correct answer by guessing. The 1PL model is sometimes called the Rasch model (Rasch, 1960), and the models are the same for most practical purposes. But there are conceptual differences, where one difference is in the case of model mismatch. Here the Rasch model suggests editing the data, to get a better model fit. Whereas, under the 1PL model approach, it is suggested to try a different model such as the 2PL or 3PL model. See e.g. von Davier (2016) for further reading on the Rasch model.

There is also possible to consider non-parametric IRT models. See e.g. Mokken (1997).

3.2 Estimation

An assessment test situation consists of the test and the examinees taking the test. The tests consist of several items, each with different item parameters. The examinee's abilities manifest through the answers to the items. Therefore, two types of parameters are estimated when fitting IRT models. One is the item parameters that consist of the discrimination (a), difficulty (b), and guessing (c) parameters. Another is the latent abilities of the examinees, now considered to be parameters instead of variables.

There exist different methods for estimating item parameters and abilities. The joint maximum likelihood, conditional maximum likelihood, marginal maximum likelihood, and Bayesian estimation are different methods for estimating item parameters. See e.g Baker and Kim (2004) for detailed descriptions. The joint maximum likelihood method estimates the item parameters and ability parameters simultaneously. The maximum likelihood, maximum a posteriori (MAP), expected a posteriori (EAP), or weighted likelihood estimation (WLE) are methods for estimating ability parameters. The methods rely on the following assumptions (Johnson, 2007).

1. The examinees are independent of each other.

2. The item responses of an examinee are independent conditional on the examinees' abilities θ_j.

Assumption 2 is the same as the local independence assumption in Section 3.1. The assumptions of independence are of importance when formulating the likelihood. Parameter estimation in IRT models is usually done by first estimating the item parameters and then estimating the ability parameters separately by using the item parameter estimates.

3.2.1 Estimation of item parameters

The joint maximum likelihood method and conditional maximum likelihood method will not be described in any further detail. The item parameter estimates are not consistent under the joint maximum likelihood method as sample size increases (Baker and Kim, 2004), and the conditional maximum likelihood method is only available for the 1PL model since it uses raw scores which is a sufficient statistic for the abilities.

We will instead consider the marginal maximum likelihood (MML) estimation procedure (Bock and Lieberman, 1970). The method removes the abilities from the likelihood by integrating them out. The MML estimation procedure therefore only treats the N examinees as observations. The examinees are assumed to be a random sample from a population where the distribution of the abilities equals a probability density function $g(\theta|\boldsymbol{\tau})$, where $\boldsymbol{\tau}$ is the vector of population parameters (Baker and Kim, 2004). Now, define $\boldsymbol{x}_j$ as the vector of item responses of the examinee with ability θ_j.

Under the assumption of conditional independence, we can write the joint probability (or likelihood) of the item response vector $\boldsymbol{x}_j$ conditional on θ_j as

$$L_j(\theta|\boldsymbol{x}_j,\boldsymbol{\beta}_i) = P(\boldsymbol{x}_j|\theta_j,\boldsymbol{\beta}_i) = \prod_{i=1}^{J} P(X_{ij} = x_{ij}|\theta_j,\boldsymbol{\beta}_i) \tag{3.1}$$

where $\boldsymbol{\beta}_i$ is the vector of item parameters. For the 3PL model, $\boldsymbol{\beta}_i = (a_i, b_i, c_i)^T$. We can now define the marginal probability of the item response vector $\boldsymbol{x}_j$

$$P(\boldsymbol{x}_j|\boldsymbol{\beta}_i) = \int_{\Theta} L_j(\theta|\boldsymbol{x}_j,\boldsymbol{\beta}_i) g(\theta|\boldsymbol{\tau}) d\theta. \tag{3.2}$$

Then the marginal log-likelihood function is

$$\log L = \sum_{j=1}^{N} P(\boldsymbol{x}_j|\boldsymbol{\beta}_i). \tag{3.3}$$

The log-likelihood in (3.3) can now be maximized with respect to the item parameters $\boldsymbol{\beta}_i$ to obtain the MML estimates. See e.g. Baker and Kim (2004) ch. 6 for more details.

Bock and Lieberman (1970) used the Newton-Raphson technique for estimating the $3n$ item parameters simultaneously. It is computationally unappealing and limited to a small number of items (Baker and Kim, 2004). The method was developed further by Bock and Aitkin (1981) by implementing the Expectation-Maximization (EM) algorithm for IRT-models. The EM algorithm is a method for computing maximum likelihood estimates where some data are assumed to be missing.

The item parameters are estimated in the presence of an unobservable random variable θ, considered to be the missing value. Let $\boldsymbol{X}$ be the observed item responses that are considered to be incomplete. Then $(\boldsymbol{X}, \theta)$ is considered to be the unobserved but incomplete data. Now let $f(\boldsymbol{X}, \theta|\boldsymbol{\beta})$ be the joint probability density function of the complete data $(\boldsymbol{X}, \theta)$, where $\boldsymbol{\beta}$ is the item parameter that is to be estimated. The EM algorithm consists of two steps.

1. E-step: The expected value of the log-likelihood of the parameters given the complete data is calculated. That is $E[\log L(\boldsymbol{X}, \theta|\boldsymbol{\beta})|\boldsymbol{X}, \boldsymbol{\beta}^p]$, where $\boldsymbol{\beta}^p$ is the item parameters in the p:th step.

2. M-step: The function that is the result from the E-step is maximized. This means that the $\boldsymbol{\beta}^{p+1}$ is chosen such that the posterior expectation is maximized.

See Baker and Kim (2004) and Bock and Aitkin (1981) fler for a more detailed description of the EM-algorithm for IRT models.

There are also Bayesian item parameter estimation methods available. See e.g. Fox (2010).

3.2.2 Estimation of abilities

The most common methods of estimation of the abilities of the examinees are the maximum likelihood estimation (MLE), the weighted likelihood estimation (WLE), expected a posteriori (EAP), and the maximum a posteriori (MAP) estimation method. The item parameters are now assumed to be estimated already, e.g. by MML estimation.

Given n items, dichotomously scored as $x_{ij} = 0, 1$, where $i = 1, \ldots, n$ are the number of items and $j = 1, \ldots, N$ are the number of examinees, the vector of responses for examinee j is denoted as $X_j = (x_{1j}, \ldots, x_{nj}|\theta_j)$. Assuming local independence, we can write the log-likelihood as

$$\log L = \sum_{i=1}^{n} (x_{ij} \log P_i(\theta_j|\boldsymbol{\beta}_i) + (1 - x_{ij}) \log(1 - P_i(\theta_j|\boldsymbol{\beta}_i))) \tag{3.4}$$

The maximum likelihood estimate of the ability θ is obtained by differentiating (3.4) and solving for 0. See Baker and Kim (2004) and Lord (1983) for more details.

A problem with MLE of the abilities θ is that they are biased outwards Lord (1983). Examinees with a perfect score on a test will have $\hat{\theta} = \infty$, and examinees that perform close to the guessing value of a multiple-choice item will have large negative values of $\hat{\theta}$. The magnitude is greater for negative values of θ. This causes problems when tests are vertically scaled (described in Section 5), where positive values from the easier test are scaled to the more biased negative values from the harder test (Warm, 1989). Lord (1983) gave an asymptotic expression for the bias of the MLE of θ.

An estimation method with small bias is the weighted likelihood estimation (WLE) method (Warm, 1989). Lord (1984) gave an expression for the bias of the Bayesian modal estimate (BME) of θ in relation to the bias of MLE. Warm (1989) used it to derive a WLE of θ that is unbiased to order n^{-1}.

A Bayesian estimator of the abilities is the expected a posteriori (EAP). It is based on

$$g(\theta_j|X_j,\boldsymbol{\beta}_i) = \frac{P(X_j|\theta_j,\boldsymbol{\beta}_i)g(\theta)}{P(X_j)} \tag{3.5}$$

where $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ is the response vector, $P(X_j) = \int_\Theta P(X_j|\theta)g(\theta)d\theta$, and $P(X_j|\theta_j,\boldsymbol{\beta_i})$ is the likelihood. The EAP-estimate is then given by

$$E(\theta_j|X_j,\boldsymbol{\beta}_i) = \frac{\int \theta_j g(\theta) \prod_{i=1}^n P_i(\theta_j)^{x_{ij}}(1-P_i(\theta_j))^{1-x_{ij}} d\theta}{\int g(\theta) \prod_{i=1}^n P_i(\theta_j)^{x_{ij}}(1-P_i(\theta_j))^{1-x_{ij}} d\theta}. \tag{3.6}$$

3.2.3 Estimation of the latent ability distribution parameters

There are also methods for estimating the latent distribution of the abilities directly. We assume here that examinees are randomly sampled from $g(\theta|\tau)$. For example, g can be a normal distribution with parameters $\tau = (\mu, \sigma^2)$.

The probability of obtaining the response pattern X_j is then $P(X_j|\theta_j,\boldsymbol{\beta}_i)$, where the item parameters $\boldsymbol{\beta}_i$ are assumed to be known. The marginal probability of X_j is then $P(X_j|\tau) = \int_\Theta P(X_j|\theta,\boldsymbol{\beta}_i)g(\theta|\tau)d\theta$.

The probability of the response matrix X for a sample of N examinees is $P(X|\tau) = \prod_{j=1}^N P(X_j|\tau)$. If we treat it as a function of τ, $L(\tau)$ can be maximized w.r.t $\tau = (\mu, \sigma^2)$. Thus the distribution parameters $\tau = (\mu, \sigma^2)$ are estimated directly without estimating the ability θ of every examinee. See Baker and Kim (2004) for a further description of the method.

3.3 Model indeterminacy

In general, IRT models are not identified. There are infinite values of item and ability parameters that will lead to the same model fit. We can observe that for the 2PL model, we have

$$P(X_{ij}=1|\theta_j)=\frac{1}{1+e^{-a_i(\theta_j-b_i)}}=\frac{1}{1+e^{-a_i((\theta_j+\delta)-(b_i+\delta))}}. \tag{3.7}$$

Equation (3.7) shows that we can shift the ability and difficulty by δ and still obtain the same probability of a correct response to the item. This is called *location* indeterminacy. If a_i is multiplied by a constant and $(\theta_j+\delta)$ and $(b_i+\delta)$ are divided by the same constant, the probability of a correct response will also be the same. It is called the *scale* indeterminacy. To solve the problem of indeterminacy, the estimates are anchored in the estimation processes. For example, when the EM-algorithm are run, the posterior ability distribution is normalized and rescaled so that $\mathrm{E}(\theta)=0$ and $\mathrm{Var}(\theta)=1$ (Baker and Kim, 2004).

3.4 Standard error and information of ability

We described how to obtain estimates of the abilities of the examinees in Section 3.2.2. It is also of interest to quantify the uncertainty of these estimates. Samejima (1977) showed that the MLE of the ability of an examinee is asymptotically normally distributed with mean θ and variance $I(\theta)^{-1}$, where $I(\theta)$ is the amount of (Fisher) information of a test that can be written

$$I(\theta)=\sum_{j=1}^{n}I_i(\theta)=\sum_{j=1}^{n}\frac{P_i'(\theta|\boldsymbol{\beta_i})^2}{P_i(\theta|\boldsymbol{\beta_i})(1-P_i(\theta|\boldsymbol{\beta_i})} \tag{3.8}$$

where $P_i'(\theta|\boldsymbol{\beta})=\frac{\partial P_i(\theta)}{\partial\theta}$. The information $I(\theta)$ is usually called the test information function (TIF) since it sums over all items of a test.

The information $I_i(\theta)$ is hence called the item information function (IIF). For the logistic IRT models, the IIF is the following (Baker and Kim, 2004)

1. 1PL: $I_i(\theta)=P_i(\theta|\boldsymbol{\beta_i})(1-P_i(\theta|\boldsymbol{\beta_i}))$
2. 2PL: $I_i(\theta)=a_i^2P_i(\theta|\boldsymbol{\beta_i})(1-P_i(\theta|\boldsymbol{\beta_i}))$
3. 3PL: $I_i(\theta)=a_i^2\frac{(1-P_i(\theta|\boldsymbol{\beta_i}))}{P_i(\theta|\boldsymbol{\beta_i})}\left(\frac{P_i(\theta|\boldsymbol{\beta_i})-c_i}{1-c_i}\right)^2$

Remark 3 *The maximum of the information for the 1PL and 2PL models are* $max[I_i(\theta)]=0.25$ *and* $max[I_i(\theta)]=0.25a_i^2$*, respectively. The maximum for the 3PL model is more complex (see e.g. Baker and Kim (2004)).*

To summarize, an estimate of a student's ability will have a smaller standard error when the item difficulties are located close to the actual ability since the standard error can be written as by Lord (1983)

$$SE^2(\hat{\theta}) = \frac{1}{I(\theta)} + o(n^{-1}), \tag{3.9}$$

which is of importance for the results in Paper II.

4. Optimal design

An optimal design is defined as a class of designs that is optimal with respect to some criterion. This statement leads us to two questions; what is a design, and what is meant by optimal?

First, we introduce the term *efficiency*, which is a concept for comparing statistical procedures, see Everitt and Skrondal (2010) p. 149. In this thesis, we will refer to the measure of the quality of an estimator. And by quality, we mean to optimize some general criterion. The variance or mean square error of an estimator are examples of such criteria.

Depending on the choice of criterion, an estimator can be efficient enough for one criterion but less efficient for another. When performing a scientific study or an experiment, the aim is usually to achieve the highest efficiency at a minimal cost. In the context of achievement tests, it means using as few examinees as possible to get an efficient enough estimator of the desired parameter. Usually, two or several procedures or estimators are compared. The ratio of the efficiencies is referred to as the *relative efficiency*.

To be able to perform an experiment or a research study, we need to choose a design. We formulate an experiment as a relationship between a set of variables X and an outcome Y (Buyske, 2005). The design of a study describes how to collect data to test the research questions, i.e. the relation between the independent variables X and the outcome Y. A study can be experimental or observational depending on if the researcher has total control of the experiment or not (Buyske, 2005).

A function or model can be fit to establish the relation between the independent variables and the outcome. The function can be linear or non-linear (see Atkinsson et al. (2007) for examples).

An optimal design can, for example, refer to estimate parameters in an optimum way according to some statistical criterion. The parameters describe the relationship between the independent variable X and the outcome Y. A classical example is the case of simple linear regression, where the slope parameter is estimated with a lower variance if the independent variables are chosen further away from the mean (Atkinsson et al. (2007) ch. 5).

Because of the relationship between the reciprocity of the variance and the Fisher information, minimizing the variance means maximizing the information. For a statistical model with more than one parameter, the optimization is done over matrices. This makes it more difficult, and several optimization criteria are available.

The inverse of the information matrix is proportional to the covariance matrix for linear models. For nonlinear models, the inverse of the information matrix is proportional to the asymptotic covariance matrix. There is also a parameter dependence for nonlinear models. Therefore, the optimization also gets more complicated in the case of nonlinear models.

4.1 Formal description

We will now formalize the optimal design theory and begin by defining the design region.

Definition 4 *The set of possible values for the independent variables X is called the design region χ.*

A design region can be either bounded or unbounded. In this thesis we assume the design region to be unidimensional, and therefore $\chi = [a,b]$ for the bounded case where $a,b \in \mathbb{R}$ and $\chi \subseteq (-\infty,\infty)$ for the unbounded case.

A design is now defined as follows (Atkinsson et al., 2007).

Definition 5 *Let $x_1,\ldots,x_n \in \chi$ be the design points and let $w_1,\ldots,w_n$, where $\sum_{i=1}^{n} w_i = 1$ and $0 \leq w_i \leq 1$, be the proportions of the observations allocated at the n distinct design points. A design is now defined as $\xi = \begin{Bmatrix} x_1 & x_2 & \ldots & x_n \\ w_1 & w_2 & \ldots & w_n \end{Bmatrix}$.*

The proportions $w_1,\ldots,w_n$ are also called design weights. If the weights are restricted to $w_i = \frac{r_i}{N}$, where r_i are the number of replications, and N is the total number of observations, the design is called an exact or discrete design. Without restrictions, the design is called an approximate or continuous design. In practice, all designs are exact (Atkinsson et al., 2007). In this thesis, we will focus on continuous designs. It can be seen that a continuous design as defined in Definition 5 is a probability measure ξ over χ and therefore $\int_{\chi} \xi(dx) = 1$.

As mentioned above, we are interested in minimizing the variance or some function of the covariance matrix, which is equal to maximizing the Fisher information or function of the information matrix. The information matrix for a continuous design ξ is denoted $I(\xi)$.

Example 2 *Consider the linear model $y = X\beta + \varepsilon$ with p parameters. Here, the $p \times p$ information matrix is $I(\xi) \propto X^{'}X$. The covariance matrix for the least-square estimators $\hat{\beta}$ is $Cov(\hat{\beta}) \propto (X^{'}X)^{-1}$.*

Remark 4 *In optimal design theory, it is common to work with the standardized information matrix $M(\xi) = \frac{1}{N}I(\xi)$ where N is the total number of observations.*

Assuming continuous designs, the objective now is to find a design that optimizes some convex function Ψ over the standardized information matrix $M(\xi)$.

Definition 6 *A design ξ^* is defined to be Ψ-optimal if $\xi^* = argmin_\xi \Psi(M(\xi))$*

There are many different optimality criteria (see Atkinsson et al. (2007)). A common criterion is the D-optimality which maximizes the determinant of the information matrix. Or equivalently minimizes

$$\Psi(M(\xi)) = \log|M^{-1}(\xi)| = -\log|M(\xi)|. \tag{4.1}$$

Another example of a criterion is the A-optimality which minimizes the trace of the inverse of the information matrix (Atkinsson et al., 2007).

Example 3 *(Continued from Example 2) In the case of a simple linear regression model, the covariance matrix of the parameter estimators is given by*

$$Cov(\hat{\beta}) = \frac{\sigma^2}{NSS_x}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{pmatrix}$$

where σ^2 is the error variance, N the sample size, and SS_x is the sum of squares. The D-criterion is now given by $\frac{\sigma^4}{NSS_x}$ and the A-criterion is given by $\frac{\sigma^2}{NSS_x}\left(\sum x_i^2 + N\right)$.

It can be seen that increasing N and SS_x will decrease the values of both criteria. Minimizing σ^2 will have the same effect. A difference between the criteria is that the D-criterion is invariant under a linear transformation of the scale of the independent variable, while the A-criterion is not.

Since derivatives of smooth functions are zero at minimum points we can use this for formulating the General Equivalence Theorem. The function Ψ depends on the measure ξ through $M(\xi)$. Let λ_x be another design that puts unit mass at x. The Frechét directional derivative (Silvey, 1980) tells how the information of the design ξ changes in the direction of the design λ_x, and can now be defined.

Definition 7 *The (Frechét) derivative of Ψ in the direction of a design λ_x, which puts unit mass at point x, is defined as*

$$F_\Psi(\xi, x) = \lim_{\alpha \to 0^+} \frac{1}{\alpha}\left(\Psi(M((1-\alpha)\xi + \alpha\lambda_x) - \Psi(M(\xi))\right).$$

The General Equivalence Theorem can now be formalized as

Theorem 4.1.1 *The following conditions on ξ^* are equivalent*

- *The design ξ^* minimizes $\Psi(M(\xi))$.*
- *The minimum over the design region χ of F_Ψ is maximized by ξ^*.*
- *The minimum over the design region χ of $F_\Psi = 0$, and occurs at the support-points of the design ξ^*.*

The theorem provides a method to check if a given design is optimal, by plotting the directional derivative. The optimal design is however not necessarily unique.

In Example 3 the optimal designs for the linear model are not dependent on the parameter values. The situation is different for non-linear models, where the information matrix depends on the unknown model parameters. The optimal design hence depends on the parameter values, and will then only be a locally optimal design. The parameter values can be given initial guesses, for example from an expert in the research area. Other strategies for handling the parameter dependence of non-linear models are the optimum in average (Bayesian), maximin, or sequential designs (see Atkinsson et al. (2007) ch. 17).

4.2 Optimal design in educational testing

There are two types of parameters in IRT, the item parameters and the ability parameter. Therefore, two kinds of optimization problems exist in educational testing (Buyske, 2005).

The first design problem describes how to select items to establish a certain level of precision of ability estimates, referred to as test design (or test assembly) (van der Linden, 2005). A test can have many purposes, which will affect the test design. A good test design is therefore crucial.

The second design problem regards sampling of test-takers to estimate the item parameter optimally. It is called item calibration (or sampling design) and is used for estimating item parameters for items to be stored in an item pool later used in test situations. For a fixed form test the aim is to construct a test such that a predetermined test information function (TIF) is satisfied. During a computerized adaptive test (CAT), a test taker (or examinee) ideally should be assigned items with difficulties close to its ability to get more precise ability estimates. It is, therefore, of great importance that the prior estimation of the item parameters is of high precision.

4.3 Test design

Birnbaum (1968) was one of the first to suggest a modern form of test construction. His approach is dependent on the test information function (TIF), which is the sum of the information for every item in a test (Equation 3.8). The idea is to construct the test to achieve the desired TIF. In the case of a certification test, the importance is to obtain high information at the ability level of the cut-off score. If the test is an aptitude test, the TIF needs to have a more uniform distribution. See Figure 4.1 for examples of test information functions.

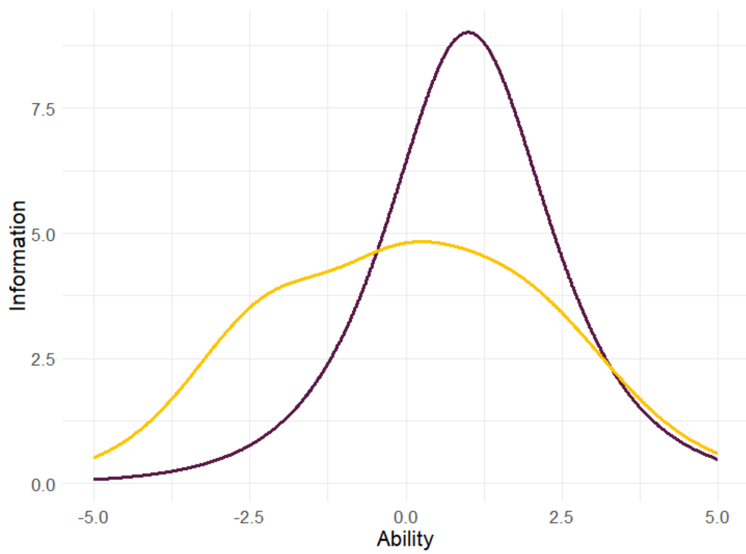


Figure 4.1: Examples of a TIF for a certification test with maximum information at ability $\theta = 1$ (purple line) and a TIF for an aptitude test (yellow line).

4.3.1 Optimal design approach

Test design aims to estimate the abilities θ of the examinees as efficiently as possible. See Berger (1998) and Buyske (2005) for a review of optimal design properties of test designs. The idea is to select items for a test such that the specified target information function of the test is fulfilled. Depending on the target, the test designs will differ.

Now assume that we would like to select a set of items with item parameters that maximize the information of the examinees' abilities that follows some distribution, e.g. a normal distribution. We let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_l)'$ be a vector of target abilities. A corresponding vector of weights $\mathbf{W} = (w_1, \ldots, w_l)'$ are introduced to describe the shape of the ability distribution. It is often common to treat $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_l)'$ as a l-dimensional parameter. Let $\mathbf{J}(\theta|\boldsymbol{\beta})$ be the $l \times l$ Fisher information matrix, which will be diagonal because of the independence assumption of IRT.

The task is now to find the items parameters $\boldsymbol{\beta}$ that maximize a function of the matrix $\mathbf{J}(\theta|\boldsymbol{\beta})$ on the ability levels of examinees where the ability distribution can be represented by $(\boldsymbol{\theta}, \mathbf{W})$. This can be formalized as maximizing $\Psi(\mathbf{J}(\theta|\boldsymbol{\beta}), \mathbf{W})$, where Ψ is a concave function. The aim is to find collection of design points $\boldsymbol{\beta}$ from the design space that maximizes Ψ.

As described in section 4.1, a common optimality criterion Ψ is the D-criterion. Since $\mathbf{J}(\theta|\boldsymbol{\beta})$ is a diagonal matrix, the D-criterion will just be the product of the diagonal elements $I(\theta_j)$.

However, since the problem is to estimate a set of unidimensional parameters it is not a multidimensional problem in the true sense. Using the D-criterion can make it more difficult to control individual objectives (van der Linden, 2005). Another common criterion, more suitable in this case, is the maximin criterion (van der Linden and Boekkooi-Timminga, 1989). It can be written as

$$\Psi(\mathbf{J}(\theta|\boldsymbol{\beta})) = \min_j (I(\theta_j)). \tag{4.2}$$

The problem can be solved using linear programming, and we can express the maximin problem as

$$\text{maximize } y$$

subject to

$$\sum_{i=1}^{n} I_i(\theta_j) t_i \geq w_j y, \text{for } j = 1, \ldots, l$$

and

$$\sum_{i=1} t_i = 1$$

and

$$t_i \geq 0, \text{for } i = 1, \ldots, n, \tag{4.3}$$

where t_i is an element of a design vector. The weights control the amount of information $I(\theta_j)$ for each ability θ level.

When constructing a fixed form test, or during the process of a CAT, the items are often taken from an item pool. An integer programming approach can be used when selecting items from an item pool (see e.g. van der Linden (2005)).

4.3.2 Design for distribution mean

We focus on the test design for estimating the growth of the mean ability of a population in this thesis. The growth can be thought of as the distance between the estimated distribution means (see Figure 4.2). Hence, it is not guaranteed

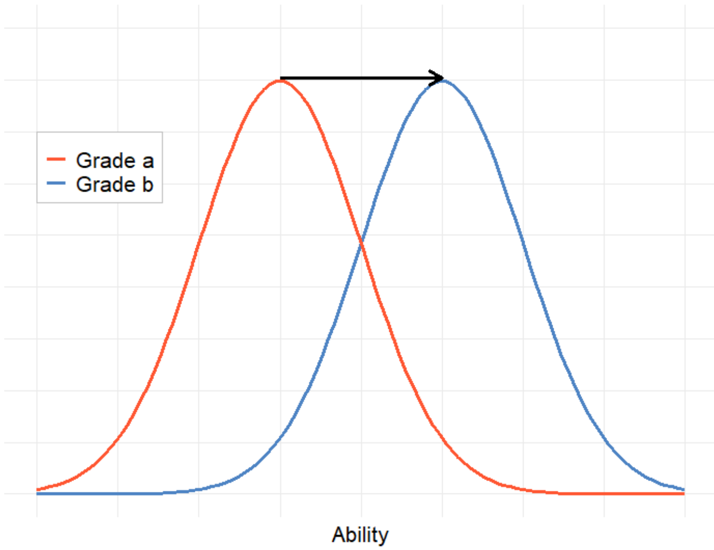


Figure 4.2: The ability distribution of the same population of examinees at different school grades, where Grade b is higher than Grade a.

that the maximin optimization criterion (or D-criterion) will give the most efficient estimation of the mean ability growth.

We now examine the objective to optimize the variance of the estimate $\hat{\mu}$ of an ability distribution. Using the asymptotic result of Equation 3.9, we can write

$$\begin{aligned} \mathrm{Var}(\hat{\mu}) &= \mathrm{Var}\left(\frac{\sum_{j=1}^{N}\hat{\theta}_j}{N}\right) = \frac{1}{N^2}\mathrm{Var}\left(\sum_{j=1}^{N}\hat{\theta}_j\right) \\ &= \frac{1}{N^2}\sum_{j=1}^{N}\left(\frac{1}{\sum_{i=1}^{n} I_i(\theta_j)}\right). \end{aligned} \tag{4.4}$$

Using an ability vector $\boldsymbol{\theta} = (\theta_1, \ldots \theta_l)'$ with corresponding vector of weights $\mathbf{W} = (w_1, \ldots w_l)'$ we can now write the optimization criterion as

$$\Psi(\mathbf{J}(\theta|\boldsymbol{\beta})) = \sum_{j=1}^{l} w_j \left(\frac{1}{\sum_{i=1}^{n} I_i(\theta_j)}\right) \tag{4.5}$$

which we will call the multi-objective weighted sum criterion. The abilities θ_j in Equation 4.5 are not the abilities of the examinees, which are unknown. The abilities θ_j are chosen by the experimenter and depend on the chosen points $1, \ldots, l$ in $\mathbb{R}$. However, the function in Equation 4.5 depends on the distribution of the examinees' abilities, which are represented by the weights.

4.4 Particle Swarm Optimization

There are several possibilities to minimize the criterion in Equation 4.5. If we would like to avoid derivatives, simulated annealing (SANN) and particle swarm optimization (PSO) are two examples of gradient-free methods for approximating a global optimum.

PSO is a computational method for optimizing by simulating social behavior (Kennedy and Eberhart, 1995). The method starts with several particles (or a swarm). The particles are randomly allocated in the search space and will move around guided by their own, and the entire particle swarms best-known positions. Improved positions of particles will inform the rest of the movements of the particle swarm. The PSO algorithm will move the particles towards the best global solution. The PSO needs values for several hyper-parameters such as percentage of particles informed, lower and upper bound, particle swarm size, and a maximum number of iterations. Clerc (2006) gives a detailed description of the optimization method.

4.5 Item calibration

The aim of item calibration is to estimate the item parameters optimally. It means to sample examinees that give the best precision in the estimates. The space of available examinees' is $\Theta = \mathbb{R}$. The task is now to determine the values of the abilities $\theta \in \Theta$ needed to estimate an item with given item parameters $\boldsymbol{\beta}$ as efficiently as possible.

First, we consider a sampling design where an examinee with any ability θ can be sampled. Using the D-optimal design as were done by Berger (1991), the 1PL and 2PL models have the following designs.

Under the 1PL model, the maximum information for b is when examinees are sampled such that $\theta = b$. For a 2PL model with item parameter $\boldsymbol{\beta}_i = (a_i, b_i)$, the optimal design puts equal weight at two design points θ_1 and θ_2 such that $p(\theta_1) = 0.176$ and $p(\theta_2) = 0.824$. This equals the design $\theta_i = b_i \pm \frac{1.543}{a_i}$ (Abdelbasit and Plackett, 1983).

Ul Hassan and Miller (2019) called such design an unrestricted design when any value of the ability is available. More formally, the unrestricted designs are represented by probability measures ξ over the design space $\chi = \Theta \times \{1,...,n\}$, $(\theta,i) \in \chi$, where θ is the ability of the examinee sampled for item i. Assume that we sample examinees with m_i ability levels $\theta_{i1}, \theta_{i2}, ..., \theta_{im_i}$ with weights $w_{i1},...,w_{im_i} \geq 0$ for all items i, $\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} = 1$. The w_{ij} is the sample weight or sample proportion for item i of examine having ability level j.

The item response model is a generalized linear model with logit link $\eta_i(\theta) = \log\left(\frac{p_i(\theta)}{1-p_i(\theta)}\right)$ that is differentiable in β_i.

The standardized information matrix of item parameters $\boldsymbol{\beta} = (\beta_1,...,\beta_n)$ is $M(\xi) = \text{diag}(M_1(\xi_1),...,M_n(\xi_n))$ with

$$M_i(\xi_i) = \sum_{j=1}^{m_i} w_{ij} v_i(\theta_{ij}) \left(\frac{\partial\eta(\theta_{ij})}{\partial\beta_i}\right)\left(\frac{\partial\eta(\theta_{ij})}{\partial\beta_i}\right)^T \tag{4.6}$$

where $v_i(\theta_{ij}) = p_i(\theta)(1-p_i(\theta))$. The non-linearity in this design implies that the optimal design depends on the item parameters. The item parameters can be given an initial value, e.g. from an expert guess.

A disadvantage with these designs is that there might not be examinees with the specified ability available for an item to be calibrated. It means that there might not be any examinee with the abilities matching the required design. Different solutions to this problem have been proposed, see e.g, van der Linden and Ren (2015) and Ren et al. (2017).

4.5.1 Restricted design

Ul Hassan and Miller (2019) proposed a solution to the problem of the sampling of examinees for a design. They developed a restricted optimal design.

Let g be a continuous density on Θ, the space of abilities of the examinees. A calibration design is described by sub-densities $h_1, h_2, \ldots, h_n \geq 0$ for each item, where $\sum_{i=1}^{n} h_i(\theta) = g(\theta)$ for all $\theta \in \Theta$.

The standardized information matrices for item i of a design h_i are then given by

$$M_i(h_i) = \int_\Theta p_i(\theta)(1-p_i(\theta))\left(\frac{\partial\eta(\theta)}{\partial\beta_i}\right)\left(\frac{\partial\eta(\theta)}{\partial\beta_i}\right)^T h_i(\theta)d\theta. \tag{4.7}$$

where the standardized information matrix of item parameters $\beta = (\beta_1,\ldots,\beta_n)$ is $M(\xi) = \text{diag}(M_1(h_1),...,M_n(h_n))$.

The D-optimality criterion for the restricted design is

$$\Psi(M(h) = \log|M^{-1}(h)| = -\log|M(h)| = -\left(\sum_{i=1}^{n}\log(|M_i(h_i)|)\right) \tag{4.8}$$

and the directional derivative of Ψ at $M(h)$ in the direction of the one-point measure $\delta_{(\theta,i)}$ in (θ,i) is

$$F_{\Psi}(h,\delta_{(\theta,i)}) = \lim_{\alpha\to 0^+} \frac{1}{\alpha}\left(\Psi(M((1-\alpha)h+\alpha\delta_{(\theta,i)}) - \Psi(M(h))\right). \tag{4.9}$$

The directional derivative describes the change in Ψ of $M(h)$ if a small number of observations in θ are added to the item i. Ul Hassan and Miller (2019) derived a new equivalence theorem for calibration of multiple items.

The theorem is used to determine if a given design h^* is optimal, by computing and plotting the n directional derivatives $F_{\Psi}(h^*,\delta_{(\theta,i)})$ for item $i = 1,\ldots,n$ over the ability space Θ. The design is optimal if the items with the smallest directional derivative under each sampling interval are allocated to the examinees with abilities in the same sampling interval.

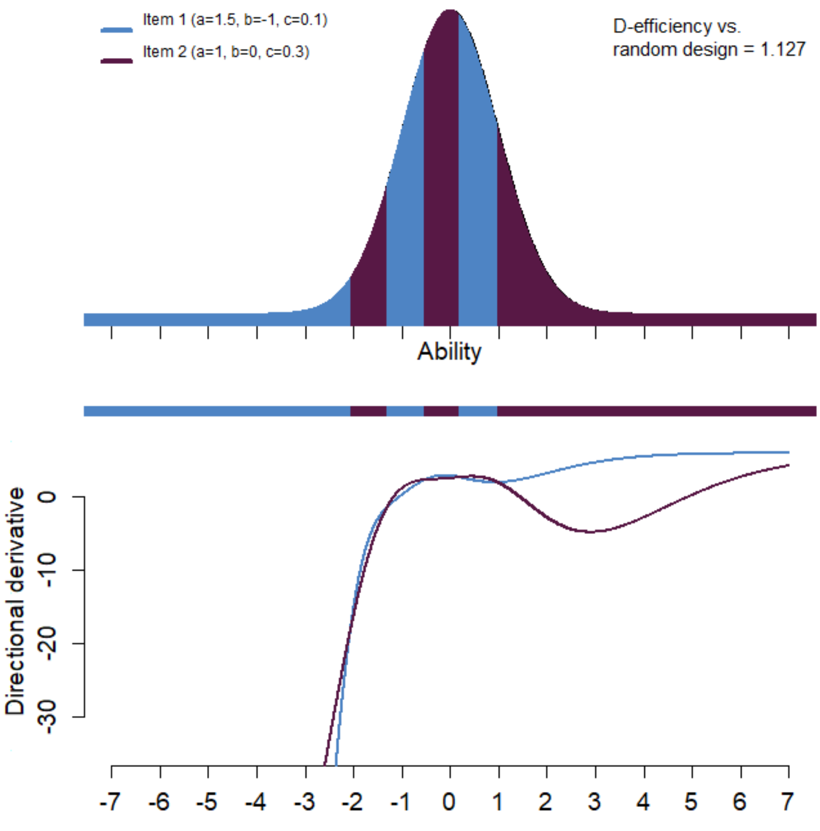


Figure 4.3: The calculated ability regions determined by the directional derivatives for two items using the exchange algorithm.

Ul Hassan and Miller (2021) developed an exchange algorithm for optimal calibration of several items. The algorithm builds on the equivalence theorem by Ul Hassan and Miller (2019) and divides the ability space Θ into small intervals. Each interval is assigned to one of the items. This means that an examinee included in the interval will calibrate the associated item. The information matrix and directional derivative are then calculated for each item. Then for each interval, it is checked whether the item for that interval has the smallest directional derivative. For the intervals with the largest violations, the item will be changed to the item with the lowest directional derivative. The algorithm will continue until some stopping criterion is met. See Ul Hassan and Miller (2021) for a detailed description of the algorithm called `OptICal`, and for a link where to find it.

After the algorithm is run the ability space Θ is divided into several intervals that are assigned to different items. It means that during an item calibration process, an examinee with ability θ will be given the item associated with that interval that includes θ. Figure 4.3 shows the ability regions and directional derivatives for an optimal design. The figure shows the plot of the regions and directional derivatives for two items that were used in the exchange algorithm. An examinee with an ability θ can only be assigned one of the items. An examinee with ability $\theta = 0$ will be assigned item 2 for calibration, and an examinee with ability $\theta = -1$ will be assigned item 1.

4.5.2 Calibration setup

The equivalence theorem for item calibration assumes that each examinee can calibrate at most one item. A blocking design can resolve this. It means that if there are several items to be calibrated, they can be divided into blocks where the `OptICal` algorithm is then run on each block separately assigning every examinee one item per block. This blocking procedure is used in Paper III where the performance of the algorithm is tested under the 3PL model.

When the exchange algorithm is run, it assumes that the examinee's abilities are the true ones. Item calibration is often conducted during a real test situation, either in a CAT format or fixed-form test. The calibration items can be included in the real test, usually in the end, or they can be given as a separate test. We called the former an integrated setup, and the latter a separated one. In both cases, the abilities of the examinees are estimated based on the operational items in the test, and the estimated ability is used in the `OptICal` algorithm. Even if the ability is estimated with high precision, there will be some uncertainty in the estimates. The uncertainty will have an impact on the efficiency of the optimization. We consider the exchange algorithm adapted to handle the uncertainty of the ability estimates in Paper IV.

4.6 Evaluation of designs

We will use several efficiency measures to compare different designs. The relative D-efficiency (RE_D) is used for comparing D-optimal designs to other designs (Atkinsson et al., 2007). It can be defined as

$$RE_D = \left(\frac{|M(\xi_r)|}{|M(\xi_o)|} \right)^{\frac{1}{P}}$$

where ξ_o is the optimal design compared to another design ξ_r. P is the number parameters in the model. A relative D-efficiency less than 1 means that the optimal design is more efficient than the relative one in terms of D-optimality.

We use two other efficiency measures in Paper III. One is the relative MSE

$$RE_{MSE} = \frac{\overline{MSE}_r}{\overline{MSE}_o}$$

where $\overline{MSE}$ is the geometric mean of the mean square error of the parameters in the IRT model. Another is the relative CC-efficiency

$$RE_{CC} = \frac{CC_r}{CC_o}$$

where CC is the squared difference between the ICC of the true item parameters and the estimated ones, summed over all items, and then summed over the examinee's abilities.

These measures are derived empirically in Paper III. The purpose of the CC-measure is to compare designs based on the ICC. Even if different design yields different estimates of the item parameters, the curves can be similar for the two designs in the region where most of the examinee's ability is present.

5. Equating and vertical scaling

To explain the concepts of equating and scaling we start with an example.

Example 4 *Imagine you and your friend are taking different versions of the same high stake test, called test Form X and Y. The test forms are similar regarding test length and content. You manage to get 70 % correct and your friend 80 %. Since you and your friend did not take the same test Form, how can we be sure who is more proficient? What if test Form X is harder than test Form Y?*

Equating is considering these questions of how to make different test scores comparable. Even if the aim is to construct test forms being similar in statistical specifications and content, it is difficult to make them equally hard. We define equating similarly as Kolen and Brennan (2014).

Definition 8 *Equating is defined as the statistical process used to adjust test scores from different test forms so that the scores can be used interchangeably.*

Interchangeably means that it should not matter if an examinee takes one or the other of two equated tests; the test score should be the same (on average). Besides the procedures of equating the tests, the content must be the same. Therefore, we make the following remark.

Remark 5 *Equating is intended to adjust for differences in difficulty between test forms, not content.*

Equating could also be considered to be the work of the test developers to construct tests that are as similar in content as possible and have the same number of items (González and Wiberg, 2017). But in this thesis, we focus on the statistical procedures for adjusting test scores to make them comparable.

The properties of equating can be summarized as follows (Kolen and Brennan, 2014).

- Symmetry Property: A function used to transform a score on Form X to Form Y is required to be the inverse of the function transforming a score on Form Y to Form X.

- Same Specifications Property: The test forms must be of the same content and statistical specifications.

- Equity Property: Examinees will have equal converted scores on Form X as on Form Y.

There are several designs for collecting data when conducting equating (González and Wiberg, 2017; Kolen and Brennan, 2014). We will only consider the non-equivalent anchor test (NEAT) design. The compared groups are considered to be random samples from different populations. The two groups are tested separately, where a part of the test consists of anchor items, the same for both tests.

There are also several statistical methods for equating (González and Wiberg, 2017; Kolen and Brennan, 2014; von Davier, 2011). We will only consider IRT equating methods in this thesis. Under the IRT model, the estimated $\hat{\theta}$ measure the examinee's ability. The process of IRT-equating will assure the ability estimates of the examinees from the different forms to be on the same scale.

Definition 9 *The transformation function mapping the IRT scale on test Form X to test form Y is defined as* $\phi : \Theta_X \to \Theta_Y$*, where* Θ *is the space of abilities.*

There exist several different methods for transforming the scores from two tests when using IRT. See e.g. Kolen and Brennan (2014), González and Wiberg (2017), and von Davier (2011) for a review of different methods. We will consider concurrent calibration (Wingersky and Lord, 1984). The parameters of both tests are estimated in a single run under concurrent calibration using some IRT software. The previously mentioned methods are referred to as separate calibration. Since no linking is needed, concurrent calibration is easier in that sense.

When the tests, that intend to measure similar constructs, differ substantially in difficulty (and likely in content), it is unlikely that the interchangeability assumption is fulfilled (Definition 8). It is then no longer possible to claim that the tests are equated. It is, therefore, more appropriate to talk about vertical scaling instead of equating. We will consider vertical scaling under the NEAT design when we are estimating mean ability growth in Paper I. When the tests are similar in content and difficulty, equating is sometimes called horizontal equating which explains why it is called vertical scaling.

Even though the equating assumptions are not full-filled, and we are considering vertical scaling there is still a matter of using concurrent or separate calibration. The important thing is to place the item- and ability parameters on the same scale for the different tests.

Separate and concurrent calibration has been compared several times, both in vertical scaling and horizontal equating situations and using real and simulated data, see e.g. Kim and Cohen (1998), Tong and Kolen (2007), Hanson and Beguin (2002), Karkee et al. (2003), Kolen and Brennan (2014), p. 444, and Lei and Zhao (2012). To summarize, it seems that concurrent calibration is better in the case of correctly specified models. Violation of the unidimensionality assumption is more problematic for the concurrent calibration method. Also, concurrent calibration has a slight advantage over separate calibration in recovering growth parameter estimates when the sample size is small and in case of a short test length.

We use the Swedish national tests (Swedish National Agency for Education, 2019) as an example in Paper I. These tests are taken place at the school level, where the number of students might be small. Therefore, we consider using the method of concurrent calibration for vertical scaling.

6. Summary of the papers

Paper I: Efficient Estimation of Mean Ability Growth Using Vertical Scaling

To compare the progression of knowledge between schools or cohorts in various subjects, it is of interest to measure students' growth of ability between different school grades. When measuring growth in student ability, precision in the estimated growth is of importance.

We compare estimation methods and test designs regarding precision and bias of the estimated growth of mean ability between two student populations that differ substantially. We compare the results using a simulation study. We assume one- and two-parameter item response models. The estimated abilities are vertically scaled using the non-equivalent anchor test design by estimating the abilities in one single run, a so-called concurrent calibration.

Different estimation methods and different choices of item difficulty parameters, called test designs, are compared to see how they affect the bias and precision of the estimated growth. We examine the effect of different test designs in this study by choosing item difficulty parameters in four different ways. We also used two different intervals of the item difficulties.

The results indicate that the expected a posteriori estimation method is preferred when estimating differences in mean ability between groups. Results also indicate that an equidistant or sorted design is preferable and that the common items do not need to have the same statistical specification as the whole tests as long as the common items have medium difficulty.

Paper II: Optimal Test Design for Estimation of Mean Ability Growth

The focus in this paper is also the mean ability growth of a population from one school grade to another. But in this paper, the optimal test design is approached more formally compared to Paper I.

We use the asymptotic expression for the mean ability growth in terms of the test information. With that expression as the criterion for optimization, we use particle swarm optimization to find the optimal design.

The design is locally optimal, meaning that the optimization function is dependent on the examinees' abilities, and therefore the values of the unknown mean ability growth. Hence, we will compare an optimum in average design to a design where the value of growth is fixed.

The main result of the paper is that we should allocate the common items in the middle of the difficulty span, with the two separate test items on different sides. The optimum in average design only improves the results very little.

Paper III: Optimal Item Calibration in the Context of the Swedish Scholastic Aptitude Test

Large-scale achievement tests require the existence of item banks with items for use in future tests. Before an item is included in the bank, its parameters need to be estimated. For the quality of the future achievement tests, it is important to perform this calibration well. It is desirable to estimate the item characteristics as efficiently as possible.

The precision of the item parameters depends on the ability of the examinee that responds to the item. The ability of the examinee is not known and must be estimated beforehand. In this paper, we address the issue of assigning items to be calibrated to the examinees based on their estimated ability.

We apply a method for calibrating items based on an optimal design algorithm. The algorithm utilizes a so-called optimal restricted design and the D-optimality criterion is used. The optimal restricted designs produce ability intervals dictating which examinees should calibrate which item. The idea here is to allocate items to examinees according to the ability intervals of the optimal design. The computations involved in the optimal design algorithm use some approximations, and the optimal design needs to be evaluated in a real testing situation.

We perform a simulation study to explore the optimal design strategy in a realistic setting and compare the results to a random design strategy, where items are instead randomly allocated to examinees. We use data from the Swedish Scholastic Aptitude Test (SweSAT) to estimate the item parameters used as a basis in the simulation study.

We use two different scenarios in the simulation study. We pre-estimate the calibration item parameters in the first scenario and then use them in the optimal design algorithm. In the other scenario, we use the true item parameters as the true ones in the algorithm. It makes it possible to study how much efficiency is lost since the optimal design solution is dependent on the item parameters. The measures used in the simulation study are developed carefully to mimic the measures used in optimal design theory. We quantify the gain from using the optimal allocation and, on average over all items, we see an improved precision of calibration. We are also able to identify for what kind of items the method works well.

Paper IV: Optimizing Calibration Designs with Uncertainty in Abilities

In computerized adaptive tests, some newly developed items are often added for calibration purposes. Methods from optimal experimental design are often used for allocating items to examinees based on their abilities. In Paper III, the examinees' abilities were both assumed to be known and pre-estimated for the allocation method. The allocation method however assumes that the abilities are the true ones.

We develop in this paper a theory for handling the uncertainty in abilities and show how to derive optimal calibration designs in this situation. We will mainly investigate the case where we can assume an asymptotic distribution of the abilities dependent on the item parameters in the operational test and the estimated abilities. We also discuss a method where we do not assume any ability distribution. Based on the estimated abilities and their uncertainty quantification, we derive optimal calibration designs.

We see that the derived optimal calibration designs are more robust if this uncertainty in abilities is acknowledged. We also implemented the method in an R package.

Future research

The setup in Paper I and II only concerns mean ability growth. If there is an interest in simultaneously measuring individual growth, there are possibilities for improvement in the methods used. The item discrimination parameters were held constant but at different levels in the optimization process in Paper II. To develop methods without these constraints would be a subject for future research. To assume different amounts of ability growth for examinees in the groups would also be of interest. Methods for handling non-normal distributions could also be considered. A further step would be to use an item bank where the item parameters are already estimated and to determine the optimal design by selecting items from the bank.

There are also possibilities to further develop the item calibration methods used in Paper III and IV. When assigning items to the examinees in Paper III, we can put further constraints on the calibration items, such as the content for example. A future task for the method handling the uncertainty in Paper IV can be to compare the method used with the naive method of using the estimated abilities as the true values. An interesting future topic would also be to develop a method that is not dependent on the normal assumption but is possible to use in an integrated setup.

7. Sammanfattning

I denna avhandling behandlar vi två områden inom kunskapsmätning. I den första delen undersöker vi hur man bäst designar två kunskapstest med gemensamma frågor så att progression i medelförmåga mellan till exempel två årskurser mäts så exakt som möjligt. I den andra delen undersöker vi hur man kalibrerar nyutvecklade frågor på ett optimalt sätt. Dessa två områden ingår i det som kallas optimal design inom kunskapsmätning.

Artikel I består av en simuleringsstudie där testfrågor allokeras på olika sätt med avseende på svårighet. I artikeln undersöker vi hur testdesign påverkar precisionen i skattning av tillväxt av förmåga i medeltal, samtidigt som vi kontrollerar för estimeringsmetod.

Vi använder oss av ett mer teoretiskt tillvägagångssätt för hur man fördelar frågor utifrån svårighet i artikel II. Här använder vi partikelsvärmoptimering på en multiobjektiv viktad summa för att bestämma en exakt utformning av de två testerna med avseende på svårighetsgrad. Resultatet är beroende av att summan som optimeras är asymptotisk lika med en funktion av testinformationen. Den allmänna slutsatsen av de båda artiklarna är att de gemensamma frågorna bör fördelas mitt i svårighetsspannet, med de två separata testens frågor fördelade på olika sidor. När skillnaden i medelförmåga mellan grupperna minskar, sammanfaller intervallen för de gemensamma och separata frågorna i högre grad.

I den andra delen undersöker vi hur man tillämpar en befintlig optimal kalibreringsmetod och algoritm med hjälp av data från högskoleprovet. Metoden vidareutvecklas sen för att kunna beakta osäkerhet i skattad kunskapsförmåga. I artikel III jämför vi tilldelning av frågor till testdeltagaren enligt den optimala kalibreringsmetoden med en slumpmässig tilldelning av frågor. Detta gör vi med en simuleringsstudie som utvärderas med hjälp av olika mått. I de flesta fall skattar den optimala kalibreringsmetoden frågorna mer effektivt, och det kan också identifieras för vilken typ av frågor metoden fungerar sämre.

Metoden som vi tillämpar i artikel III förutsätter att de skattade testdeltagarnas förmåga är de sanna. I artikel IV utvecklar vi därför kalibreringsmetoden för att hantera osäkerhet i skattningarna av testdeltagarnas förmågor. Vi undersöker ett asymptotiskt resultat och jämför det med fallet med där testdeltagarnas förmågor är kända. Resultatet för den optimala kalibreringsmetoden som använder skattningar närmar sig då den optimala metoden som använder sanna förmågor, när mer information från ett operativt test är tillgängligt.

References

Abdelbasit, K. M. and Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association*, 78(381):90–98. 24

Atkinsson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press, Oxford. 17, 18, 19, 20, 28

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques. 2nd Edition*. CRC Press, Boca Raton, Florida. 7, 9, 10, 11, 12, 13, 14

Berger, M. P. F. (1991). On the efficiency of irt models when applied to different sampling designs. *Applied Psychological Measurement*, 15(3):293–306. 24

Berger, M. P. F. (1998). Optimal design of tests with dichotomous and polytomous items. *Applied Psychological Measurement*, 3(22):248–258. 21

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability. In: Statistical theories of mental test scores (eds. F. M. Lord and M. R. Novick)*. Addison-Wesley, Reading, Massachussetts. 21

Bock, D. R. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46:443–459. 12

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35:179–197. 11

Buyske, S. (2005). Optimal design in educational testing. In Berger, M. P. F. and Wong, W. K., editors, *Applied Optimal Design*, pages 1–19. John Wiley and Sons, Ltd. 17, 20, 21

Clerc, M. (2006). *Particle Swarm Optimization*. John Wiley and Sons, Ltd. 24

Everitt, B. S. and Skrondal, A. (2010). *The Cambridge Dictionary of Statistics, Fourth Edition*. Cambridge University Press, The Edinburh Building, Cambridge CB2 8RU, UK. 17

Fox, J.-P. (2010). *Bayesian Item Response Modeling - Theory and Applications*. Springer, New York. 5, 12

González, J. and Wiberg, M. (2017). *Applying Test Equating Methods Using R*. Springer, Switzerland. 29, 30

Hanson, B. A. and Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1):3–24. 31

International Study Center, Boston College (2017). TIMSS and PIRLS. `https://timssandpirls.bc.edu/timss2015/international-database/assessment-items.html` (last access August 26, 2020). 1

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software, Articles*, 20(10):1–24. 8, 10

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., and Haug, C. (2003). Separate versus concurrent calibration methods in vertical scaling. *Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22- 24, 2003)*. 31

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4. 24

Kim, S.-H. and Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22(2):131–141. 31

Kolen, M. J. and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. Springer, New York. 29, 30, 31

Lei, P. W. and Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement*, 36(1):21–39. 31

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their paralell-forms reliability. *Psychometrika*, 48(2):233–245. 13, 15

Lord, F. M. (1984). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23(2):157–162. 13

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In van der Linden, W. J. and Hambleton, R. K., editors, *Handbook of Modern Item Response Theory*. Springer, New York, NY. 10

Nering, M. L. and Ostini, R., editors (2010). *Handbook of Polytomous Item Response Theory Models : Developments and Applications*. Routledge, 270 Madison Avenue, New York. 7

OECD (2019). PISA 2018. `https://www.oecd-ilibrary.org/education/pisa-2018-results-volume-i_5f07c754-en` (last access August 31, 2021). 1

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, Denmark. 10

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer, New York, NY. 7

Ren, H., van der Linden, W. J., and Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika*, 82:498–522. 25

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement*, 1(2):233–247. 14

Silvey, S. D. (1980). *Optimal Design*. Chapman and Hall. 19

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. Chapman and Hall/CRC, Boca Raton, Florida. 5

Swedish National Agency for Education (2019). Nationella prov – obligatoriska prov i skolan. `https://www.skolverket.se/for-dig-som-ar.../elev-eller-foralder/betyg-och-nationella-prov/nationella-prov---obligatoriska-prov-i-skolan` (last access August 26, 2020). 1, 31

Tong, Y. and Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2):227–253. 31

Trafikteori Online (2021). When you want to obtain a Swedish Driving License. http://www.drivinglicense.se/theorytest.php (last access September 2, 2021). 1

Ul Hassan, M. and Miller, F. (2019). Optimal item calibration for computerized achievement tests. *Psychometrika*, 84:1101–1128. 2, 24, 25, 26, 27

Ul Hassan, M. and Miller, F. (2021). An exchange algorithm for optimal calibration of items in computerized achievement tests. *Computational Statistics and Data Analysis*, 157:107–177. 2, 26, 27

Umeå University (2021). The Swedish Scholastic Aptitude Test - SweSAT. https://www.umu.se/en/department-of-applied-educational-science/swesat/ (last access May 17, 2021). 1, 2

van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. Springer, New York. 20, 22

van der Linden, W. J. and Boekkooi-Timminga, E. (1989). Binary programming and test design. *Psychometrika*, 53(2):237–247. 22

van der Linden, W. J. and Hambleton, R. K., editors (1997). *Handbook of Modern Item Response Theory*. Springer, New York, NY. 7

van der Linden, W. J. and Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, 80:263–288. 25

von Davier, A. A. (2011). *Statistical Models for Test Equating, Scaling, and Linking*. Springer, New York. 30

von Davier, M. (2016). Rasch model. In van der Linden, W. J., editor, *Handbook of Item Response Theory*, chapter 3. CRC Press. 10

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3):427–450. 13

Wingersky, M. S. and Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, 8:453–461. 30