Extensions of the Kernel Method of Test Score Equating

Gabriel Wallin

Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå 2019
Contents

List of papers iv

Abstract v

Populärvetenskaplig sammanfattning vi

Acknowledgements viii

1 Introduction 1

2 SweSAT 2

3 Test score data 3
  3.1 Notation of score variables ......................... 3
  3.2 Data collection designs ............................. 4
  3.3 Propensity scores ................................. 5

4 Observed-score test equating 7
  4.1 Requirements for equating .......................... 7
  4.2 Definitions of equivalent scores ................... 8
  4.3 The kernel method of test score equating ........... 9
    4.3.1 Presmoothing ..................................... 9
    4.3.2 Calculating the score probabilities ............ 12
    4.3.3 Continuization ................................... 14
    4.3.4 Equating ......................................... 15
    4.3.5 Evaluation of the equating function ............ 17

5 Summary of papers 19
  5.1 Paper I ........................................... 19
  5.2 Paper II .......................................... 20
  5.3 Paper III ......................................... 20
  5.4 Paper IV .......................................... 21

6 Final remarks and further research 21

Papers I-IV
List of papers

The thesis is based on the following papers:


Abstract

This thesis makes contributions within the area of test score equating and specifically kernel equating. The first paper of this thesis studies the estimation of the test score distributions needed in kernel equating. There are currently two families of models implemented within kernel equating for this purpose, namely log-linear models and item response theory models. The impact of model selection criteria for these models on the equated scores are studied using both empirical and simulated data.

The second paper focuses on the continuization of the estimated score distributions, where different bandwidth selection methods are studied. The study considers multiple data collection designs, sample sizes and score distributions and investigates how large impact the bandwidth selection has on the equated scores.

When the test groups differ in their ability distributions it is necessary to adjust for such differences to make fair comparisons possible. The most common way of adjusting for ability imbalance is by using items that are common for both test forms, known as anchor items. If no such items are available, it has been suggested to use background information about the test-takers instead. However, when the covariate vector of background information increases, the combination of covariates with no observations tends to increase as well. The third paper of this thesis therefore suggests to transform the covariate vector into a scalar propensity score. Two equating estimators are suggested under this setting and the standard error of equating (SEE) for both estimators are given.

The SEE for kernel equating has previously been derived using the delta method. The fourth paper revisits the Bahadur representation of sample quantiles to derive the SEE. Both methods of calculating the SEE are compared, and it is shown that they are equivalent for all common data collection designs when the terms of the Bahadur SEE are estimated using Taylor expansions. An implementation of an alternative estimator of the Bahadur SEE for which the equivalence result does not hold is also included to illustrate when the two methods differ.

KEYWORDS: Test equating, Nonequivalent groups, Standard error of equating, bandwidth selection, log-linear models, item response theory.
Populärvetenskaplig sammanfattning


Provet som mätinstrument är problematiskt eftersom det i regel byts ut i sin helhet från ett provtillsfälle till ett annat. Det är därför eftersträvansvärt att göra prov som är lika svåra. Detta är dock mycket svårt att till fullo uppnå. Att t.ex. få 70 % rätt på en version av ett prov behöver däremot inte nödvändigtvis indikera samma kunskapsnivå som att få 70 % rätt på en annan version av provet. För att rättvisa jämförelser mellan olika provtagargrupper ska vara möjliga måste provpoängen därför justeras i enlighet med den skillnad i svårighetsgrad som proven har. Denna typ av poängjustering kallas för ekvivalering och utförs idag på mer eller mindre samtliga storskaliga kunskapsmätningar över hela världen.

Denna avhandling bidrar med fyra artiklar där ekvivalering studeras och utvärderas. En viktig del av en ekvivalering är att skatta sannolikheterna för respektive provpoäng. I den första och andra artikeln så studeras hur detta görs på bästa möjliga sätt. Olika metoder och modeller studeras och utvärderas bland annat med hjälp av datorsimuleringar.

Den tredje artikeln studerar den vanligt förekommande situationen att de två grupper av individer som ska jämföras skiljer sig åt gällande kunskapsnivå. De har dessutom skrivit olika svåra prov vilket gör det svårt att avgöra om skillnaden i provresultat mellan grupperna beror på att den ena gruppen har mer kunskap än den andra, eller om skillnaden
enbart beror på att det ena provet var svårare än det andra. I artikeln så föreslås metoder för att kunna separera dessa två effekter med hjälp av bakgrundsinformation om provtagarna, som ålder och kön.

När ekvivaleringen har utförts så återstår att utvärdera den. En vanlig utvärderingsmetod är att ange osäkerheten i ekvivaleringen i siffror genom att beräkna standardavvikelsen för ekvivaleringsmetoden. Den fjärde artikeln i denna avhandling studerar två olika sätt att utföra denna beräkning och visar ett fall där dessa två metoder sammanfaller.
Acknowledgements

It was probably not the smartest thing of me to let this part of the thesis be the last one to write since it most likely will be the only part that people will read, with a few exceptions. On the other hand, it seemed a bit too confident to write this part during the first year as a PhD student. There is however no doubt that this thesis would not have been possible without the help of a few people. I would like to start by thanking my main supervisor Marie Wiberg. I am so grateful for your always optimistic mindset, structured way of working and generosity with your time. I have learned so much more about research than what is covered by this thesis thanks to you and I cannot imagine having a better supervisor. I also want to thank my co-supervisor Jenny Häggström. Your knowledge, support and ability to ask lots of relevant questions when we have tried to disentangle the sometimes confusing world of simulation studies in equating research have been crucial.

I have also been fortunate to collaborate with researchers outside of the department. I therefore would like to thank Jorge González. It has been a true pleasure working with you, much thanks to your enthusiasm for statistics. I also would like to thank Irini Moustaki and Yunxiao Chen who invited me to visit the Department of Statistics at London School of Economics and Political Science. I learned a lot during my stay and I hope for future collaborations. Many thanks as well to the PhD students that gave me such a warm welcome in the office. Thank you for letting me introduce some authentic Swedish concepts like fika, semla and champagne Thursdays!

During these years I have shared office with a number of people, and I am so grateful to you all. Thank you Philip, Anita and Inga who made the first period as a PhD student such a nice experience. I’m proud to tell you that I never had to use the napkins intended for “PhD tears” that you gave me when I started. It’s much thanks to you! A special thanks to Tanya for joining me to Iksu so many times, and to Anita for all the support, especially the last hectic months. It has made a big difference, true story! A big thank you to all former and present colleagues at the Department of Statistics as well! It has been a pleasure being a student and co-worker at this department.

Thanks to my friends Waldir and Denise, and to all members of the extended pancake family! A special thanks to Wilmer, Carl Larsson,
Bore, John (Fristad!), Robin and Hanna. Thank you all for making my life so much more fun, and to some of you for always giving me a stomach ache on Thursday afternoons!

A big thank you to my family for always supporting me. A special thanks to my grandfather who taught me how to swim, drive a car and basically all the math I know. These qualities have all, in one way or another, come to good use during these years! Thanks mom for always encouraging and supporting me in everything I have ever attempted.


Umeå, December 2019
Gabriel Wallin
1 Introduction

Through most branches of science, the comparability of measurements generated by a common phenomenon is often a key objective. Within educational measurement, one of the main tasks is to compare test-takers in terms of a latent construct. When the construct is some sort of ability, the test score from an assessment test is often used as a proxy for the latent variable. The test score from a large-scale assessment is commonly used as a selection instrument in the application process for higher education. It thus has the possibility to change a person’s life trajectory. To justify this use of test scores, educational testing programs strive to produce assessments that are characterized by fairness and high validity. There are however several factors that can challenge the quality of an assessment. Items that do not effectively discriminate between high- and low-achievers, items that function differently in subgroups of the population and cheating are to some extent present in most tests. Different measures are therefore made to reduce the impact of such disturbances. Some of these actions however cause other problems. One such example is test form difficulty, which necessarily differs between test administrations since items are usually unique for each administration. Such differences however make it impossible to compare test-takers without making any adjustment of the test scores. Within educational measurement, such score adjustment is a statistical process known as test score equating.

This thesis makes contributions to the equating framework of kernel equating. In Paper I of the thesis, the impact of model choice criterion is studied for log-linear and item response theory (IRT) models when the purpose is to estimate the score distributions to do equating. The score distributions need to be continuous to make equating possible. Since test scores are most often discrete, a kernel function is utilized to make continuous approximations of the estimated, discrete score distributions. For this purpose, a smoothing parameter, called the bandwidth, needs to be selected. In Paper II, the bandwidth selection is studied. Specifically, all current bandwidth selection methods together with two new methods that implement leave-one-out cross-validation are evaluated.

When the test forms to be equated are confounded by test group ability, there is a need to adjust for such differences. The gold standard is to utilize a set of common items, known as anchor items. However,
not all testing programs have the possibility to administer anchor items. Paper III studies the possibility to adjust for this confounding by using background information about the test-takers instead of anchor items. The idea is that these covariates can be used as a proxy for ability and therefore adjust potential differences. Due to dimensionality problems, it is not practical to use more than one or two covariates before the number of empty cells proliferates. Paper III shows how to gather the covariates in a scalar function of the covariate vector, known as the propensity score. This way of handling ability confounding is evaluated in a simulation study and using real test data.

As with all statistical estimators, the equating function exhibits sampling variability. The standard error of equating (SEE) is one of the most common evaluation measures of an equating function. For most equating estimators, the SEE is derived using the delta method. In Liou and Cheng (1995) and Liou, Cheng and Johnson (1997) however, the SEE is derived using a different asymptotic result known as the Bahadur representation of sample quantiles. In Paper IV, the Bahadur result for the SEE is generalized to fit all common data collection designs and kernel functions. It is furthermore shown under what circumstances the SEE resulting from using the Bahadur result is equivalent to the SEE resulting from using the delta method.

This thesis has the following structure: First the SweSAT is presented, followed by an overview of the data collection designs that this thesis considers. Next, the kernel equating framework is described. Papers I-IV are then summarized, and the thesis ends with concluding remarks and an outlook for future research.

2 SweSAT

The Swedish Scholastic Assessment Test (SweSAT) is a paper-and-pencil test that prospective university students have the possibility of taking as part of their application for higher education. It has been administered since 1977 and today, the Swedish educational system allows students to apply for university programs using both their high school grades and their results on the SweSAT. By taking the test, test-takers thus have the possibility to compete for a spot on an educational program in two selection groups. The test is administered twice a year and the test
result is valid for five years. There is no upper limit on how many times one is allowed to take the test, and only the best test result is counted (Lyrén and Hambleton, 2011).

It takes approximately two years to produce a new SweSAT administration. The test consists of one quantitative section and one verbal section, each consisting of 80 items. The test measures general abilities like reading comprehension and quantitative ability that are associated with a good performance at the university level (Lyrén and Hambleton, 2011).

SweSAT is equated in two steps, treating the two sections separately. For each section, the total group of test-takers in each administration is divided into two reference groups. The first reference group is selected such that the distributions of age, sex and educational background match previous test administrations. The second reference group contains students between an age of 18 to 19 years that are attending an academically oriented program. Equating is thereafter separately conducted using equipercentile equating, which will be described in Section 4, for both reference groups and for the total group of test-takers. The final equating function is then the result of a weighted average of the three separate equatings, where the total group of test-takers is given the largest weight (Lyrén and Hambleton, 2011). The equated scores are lastly translated to a scale score ranging from 0.0 to 2.0, with increments of 0.1. It is only the scale score that is reported to the test-takers. In this thesis, papers I and II use the spring administration of 2012 and the autumn administration of 2011, and paper III uses the two administrations of 2015. For these three papers, the quantitative section of the test was used in the respective empirical studies.

3 Test score data

3.1 Notation of score variables

There are a number of ways to separate the effect of test-taker ability from that of test form difficulty. The method chosen is largely controlled by how the data is collected and generally there are two methods implemented by testing programs, one that makes use of common test-takers, and one that makes use of common items (von Davier, 2013). In this section, these approaches are discussed for the data collection designs.
that this thesis mainly considers.

Firstly, let $X$ denote the test score from a new test form $X$, and let $Y$ denote the test score from an existing, or old, test form $Y$. The test-takers that were administered test form $X$ are considered to be a random sample from a population of test-takers denoted by $P$, and the test-takers that were administered test form $Y$ are assumed to be a random sample from a population denoted by $Q$. Populations $P$ and $Q$ are potentially different, but can also be treated as equal depending on the data collection design. Because of the random sampling of test-takers, the scores $X$ and $Y$ are treated as random variables. The realizations of $X$ and $Y$ are denoted by $x_j$ and $y_k$, respectively, where $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Throughout this thesis, it is assumed that the total score is the sum of dichotomously scored items, so that $x_1 = 0, \ldots, x_J = J - 1$ and $y_1 = 0, \ldots, y_K = K - 1$. Test form $X$ thus consists of $x_J$ items and test form $Y$ consists of $y_K$ items.

3.2 Data collection designs

In the equivalent groups (EG) design, the two test groups to be equated are treated as random samples from the same underlying population of potential test-takers (Kolen and Brennan, 2014). It is thus assumed that the test groups are only randomly different in terms of the latent ability that the test is supposed to measure. The test groups are administered different test forms and the task is to adjust the scores for differences in test form difficulty. The test groups thus contain common test-takers, common in terms of the latent trait. This design is represented in Table 1.

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$P$</td>
<td>2</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

If the test groups differ in their respective ability distributions, the gold standard is to use common items, called anchor items, to adjust for such differences. The anchor items are intended to measure the same underlying construct as the main test, and they are therefore often strongly correlated with the other test items (von Davier et al., 2004).
In this design, referred to as the non-equivalent groups with anchor test (NEAT) design, the test groups are viewed as random samples from different populations, $P$ and $Q$. The test group from population $P$ is administered test form X and the test group from population $Q$ is administered test form Y. All test-takers in both groups have a test score $A$ on the anchor test, with a realization denoted by $a_l$, $l = 1, \ldots, L$. This design is summarized in Table 2.

Table 2: The non-equivalent groups with anchor test design.

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>X</th>
<th>Y</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>$Q$</td>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

If the test groups are non-equivalent and no anchor items are available, it has been suggested to use background information about the test-takers to adjust for the ability difference. This is known as the non-equivalent groups with covariates design (NEC; Wiberg and Bränberg 2015). Letting $D$ denote a vector of covariates measured on all test-takers in both test groups, the non-equivalent groups with covariates design is summarized in Table 3.

Table 3: The non-equivalent groups with covariates design.

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>X</th>
<th>Y</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>$Q$</td>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Propensity scores

In the NEC design, the purpose is to balance the test groups in terms of the covariate vector. The underlying assumption is that the covariates are related to the latent ability, so if the groups are balanced on the covariate vector they should be approximately balanced on the latent ability. It is however not clear how to best utilize the information in the covariate vector to achieve this purpose. There are also practical challenges to handle more than a few covariates. If the number of covariates increases, the number of combinations for which there are no test-takers will increase as well. To include more information, it is therefore a rea-
sonable idea to investigate dimension-reducing techniques. One common such technique is to use the so-called propensity score.

Let \( Z \) denote a binary random variable equal to 1 if a test-taker is administered the old test form \( X \), and 0 if administered the new test form \( Y \). Borrowing terminology from the field of causal inference, the case of \( Z = 1 \) is referred to as an assignment to the treatment group, and that of \( Z = 0 \) as an assignment to the control group (Rosenbaum and Rubin, 1983). Treatment is thus defined as a test-taker taking the old test form, but it is possible to define \( Z \) in the opposite manner. The propensity score equals the conditional probability of getting the treatment given the covariate vector, i.e.,

\[
e(D) = \Pr(Z = 1|D).
\]

The propensity score therefore is a scalar function of the covariate vector, meaning that all available information about the test-takers can be incorporated into the covariate vector without increasing the dimension.

The propensity score is furthermore a balancing score, \( b(D) \), meaning that the conditional distribution of \( D \) given \( b(D) \) is the same for the treatment and control group,

\[
D \perp Z|b(D).
\]

Rosenbaum and Rubin (1983) showed that \( e(D) \) is the coarsest balancing score. It is therefore sufficient to control for \( e(D) \) to create balance in the test groups with respect to the covariate vector \( D \). If \( D \) includes all covariates that affect both \((X,Y)\) and \( Z \), i.e. all confounders, the functional relationship between \( X \) and \( Y \) is possible to identify with data.

The propensity score is unknown and needs to be estimated. One common estimation technique is to use logistic regression. It should be emphasized that the purpose set out for the propensity score is covariate balance, so usual goodness-of-fit statistics are not the most suitable measures to use. Instead, the balancing property should be guiding the model specification: If there is dependence between \( D \) and \( Z \) conditional on the estimated propensity score, it would be an indication of model misspecification. A new model can then be specified, and the covariate balance can be checked again.

There are a number of ways that propensity scores can be used and this is to some part driven by the application. In Paper III of this thesis,
the approach taken by Rosenbaum and Rubin (1984) is used to create strata for which covariate balance holds in the test groups. Using the propensity score as a proxy for ability, the ability imbalance in each strata is controlled for and two estimators of the equating transformation are derived.

4 Observed-score test equating

4.1 Requirements for equating

The target parameter in this thesis is a function that relates the test scores from the new test form $X$ to the old test form $Y$. However, not all functions that map the scores from the sample space of $X$, denoted $\mathcal{X}$, to the sample space of $Y$, denoted $\mathcal{Y}$, are equating functions. Five requirements are often presented to define what constitutes equating when test scores from different administrations are to be related to each other. These requirements are:

a. *Equal constructs.* The test forms measure the same underlying construct.

b. *Equal reliability.* The test forms are equally reliable.

c. *Symmetry.* There is no practical difference in equating the old test form to the new or vice versa.

d. *Equity.* After equating, it should not matter which test form the test-taker took.

e. *Population invariance.* The estimated equating function is not affected by the choice of sub-population for which it has been calculated.

As pointed out in Dorans and Holland (2000), these requirements should be viewed as guidelines rather than testable assumptions.

Following the five requirements above, there are a number of such functions suggested by previous research and implemented by testing programs. These equating functions are often classified as either traditional or modern. The main difference between the methods is in how they define equivalent scores. To the class of traditional methods belong
e.g. mean, linear and equipercentile equating, and to modern methods e.g. local equating, IRT equating and kernel equating (González and Wiberg, 2017).

4.2 Definitions of equivalent scores

The mean equating function defines two scores $x \in X$ and $y \in Y$ as equivalent if they are equally distant to their respective means:

$$x - \mu_X = y - \mu_Y,$$

As the mean equating function can only account for constant differences between the test forms, it is mostly used for illustrative purposes (González and Wiberg, 2017).

The linear equating function defines scores $x \in X$ and $y \in Y$ as equivalent if they are equally distant to their respective means, measured in standard deviations:

$$\frac{x - \mu_X}{\sigma_X} = \frac{y - \mu_Y}{\sigma_Y}.$$

However, the linear equating function is insufficient for tests where for example the old test form is more difficult at high and low scores, but less difficult at the middle scores. In other words, when the relationship between $X$ and $Y$ is non-linear.

The most flexible equating transformation defines two scores $x \in X$ and $y \in Y$ as equivalent if

$$G_Y(y) = F_X(x),$$

where $G_Y(\cdot)$ denotes the cumulative distribution function (CDF) of $Y$ and $F_X(\cdot)$ denotes the CDF of $X$. Expressing this equivalence in terms of a score $y$ yields a general formula for comparing the distributions of two random variables:

$$y = \varphi(x) = G_Y^{-1}(F_X(x)) \quad (1)$$

(Wilk and Gnanadesikan, 1968). In the equating literature, $\varphi(x)$ is known as the equipercentile transformation as it identifies scores from two test forms as equivalent if they have the same percentile rank in their respective distributions (González and Wiberg, 2017). The equipercentile transformation allows for a non-linear relationship between $X$
and $Y$, but requires that the score variables $X$ and $Y$ are continuous. If not, there might be several scores on the new test form that equates to the same score on the old test form. Traditionally, this has been solved by linearly interpolating the score points, creating piecewise continuous functions. In this thesis the equipercentile transformation will be utilized, but instead of linear interpolation, kernel smoothing techniques (Silverman, 1986) will be used to approximate the discrete score distributions with continuous functions.

4.3 The kernel method of test score equating

The kernel method of test score equating, or simply kernel equating, is an equating framework consisting of five steps: Presmoothing, estimating the score probabilities, continuizing the estimated score CDFs, equating, and evaluating the equating function (von Davier et al., 2004; González and Wiberg, 2017). In what follows, each step of the process will be described.

4.3.1 Presmoothing

The equipercentile transformation is defined as a functional composition of the score CDFs, meaning that it is dependent on the score probabilities (von Davier, 2013). Let the vector of score probabilities for the target population $T$ on which the equating is computed be denoted by $\mathbf{r}$ and $\mathbf{s}$,

$$\mathbf{r} = (r_1, ..., r_J)\top \quad \text{and} \quad \mathbf{s} = (s_1, ..., s_K)\top,$$

where $r_j = \Pr(X = x_j|T)$ and $s_k = \Pr(Y = y_k|T)$. Although it is possible to use the observed relative frequencies as estimates of $\mathbf{r}$ and $\mathbf{s}$, it has been shown that presmoothing the score distributions by fitting a statistical model to the observed data reduces sampling variability and results in a more stable equating estimator (Kolen and Brennan, 2014). Within the kernel equating framework there have been two presmoothing model suggestions, log-linear models (Holland and Thayer, 1989, 2000) and IRT models (Andersson and Wiberg, 2017), where the properties of the former have been widely studied in the equating context, see for example Moses and Holland (2009), Moses and von Davier (2011), and Liu and Kolen (2019).
To define the log-linear model, let $n_j$ denote the number of test-takers scoring $X = x_j$ and $m_k$ denote the number of test-takers scoring $Y = y_k$. Further let $\mathbf{p}$ and $\mathbf{q}$ denote the probability vectors of $\mathbf{n} = (n_1, \ldots, n_J)^\top$ and $\mathbf{m} = (m_1, \ldots, m_K)^\top$, respectively. It is assumed that the vectors $\mathbf{n}$ and $\mathbf{m}$ are independent and follow multinomial distributions $\text{MN}(N, \mathbf{p})$ and $\text{MN}(M, \mathbf{q})$, respectively, where $\sum_j n_j = N$ and $\sum_k m_k = M$. Under these assumptions, it is possible to represent the $j$th entry $p_j$ in $\mathbf{p}$ as a function of the test scores,

$$\log(p_j) = \beta_0 + \sum_{i=1}^I \beta_i x_{ij},$$

(2)

where $\beta_0$ is a normalizing constant, the $\beta_i$:s are regression coefficients to be estimated and $x_{ij}$ is a function of the test scores. When the regression coefficients are estimated using maximum likelihood the moments of the estimated score distributions equal those of the observed distributions,

$$\sum_j \frac{n_j}{N} x_{ij} = \sum_j \hat{p}_j x_{ij} \quad \forall I,$$

and the corresponding identity holds for the $Y$ score distribution. The moment-matching property means that the log-linear model in Equation 2 preserves $I$ moments of the observed test score distribution (Moses and Holland, 2010).

In this thesis, log-linear models are frequently used to estimate bivariate distributions. Under similar assumptions as made for the univariate case, and with $p_{jl} = \Pr(X = x_j, A = a_l)$, the log-linear model for the bivariate distribution of test score $X$ and anchor score $A$ can be expressed as

$$\log(p_{jl}) = \beta_0 + \sum_{i=1}^I \beta_{x,i} x_{ij} + \sum_{h=1}^H \beta_{a,h} a_k^h + \sum_{d=1}^D \sum_{e=1}^E \beta_{x,a,de} x_{jd} a_k^e.$$

The moment-matching property holds for the bivariate log-linear model as well when the model parameters are estimated using maximum likelihood. This means that $I$ sample moments of the marginal distribution of $X$ are preserved, $H$ sample moments of the marginal distribution of $A$ are preserved, and $D$ and $E$ determine the number of cross-moments that are preserved in the joint distribution of $(X, A)$. 

10
Another way of estimating the score probabilities is through the use of an IRT model. This approach was implemented to the kernel equating framework by Andersson and Wiberg (2017). IRT is used to model the probability of correctly answering an item and the underlying assumption is that such a probability is a function of a latent variable $\theta$ and a set of item parameters that characterizes the function (Hambleton and Swaminathan, 1985). A common IRT model is the three-parameter logistic model, which models the probability of a randomly selected test-taker answering item $j^* \in \{1, 2, ..., x_J\}$ correctly (Lord, 1980). For this purpose, let $X_{ij^*} = \{0, 1\}$ be the test score on item $j^*$ by test-taker $i$, let $\theta_i \in (-\infty, \infty)$ denote the latent ability of test-taker $i$, let $\alpha_{j^*} \in [0, \infty)$ denote the discrimination of item $j^*$, $b_{j^*} \in (-\infty, \infty)$ denote the difficulty level of item $j^*$ and $c_{j^*} \in [0, 1]$ denote the lower asymptote (or guessing parameter) of item $j^*$. Using the three-parameter logistic model, the probability that test-taker $i$ answers item $j^*$ correctly equals

$$\Pr(X_{ij^*} = 1|\alpha_{j^*}, b_{j^*}, c_{j^*}; \theta_i) = p_{ij^*} = c_{j^*} + \frac{1 - c_{j^*}}{1 + \exp(\alpha_{j^*}(\theta_i - b_{j^*}))}$$

Setting $c_{j^*} = 0$ in Equation 3 yields the two-parameter logistic model, and if additionally $\alpha_{j^*} = 1$, the one-parameter logistic (or Rasch) model is obtained. When the item parameters of the IRT model are estimated using maximum likelihood (Bock and Aitkin, 1981) they are asymptotically multivariate normally distributed (Ogasawara, 2009), which can be exploited when deriving the SEE.

When the class of presmoothing model has been selected there are two categories of model selection statistics, one based on significance testing strategies, and one based on parsimonious model selection criteria. To the former belongs e.g. the likelihood ratio chi-square statistic, and to the latter belong e.g. the Akaike information criterion (AIC; Akaike 1974) and the Bayesian information criterion (BIC; Schwarz 1978). All three of these measures are commonly used for both log-linear and IRT models and they have the possibility of selecting different parameterizations of the log-linear and IRT model, respectively.

To illustrate the presmoothing step, a sample from the spring administration of the SweSAT from 2012 and a sample from the autumn administration of the SweSAT from 2011 are used. In Figure 1, the observed-score frequencies are displayed for both administra-
tions, adopting an EG design. In the upper panel, the sample from the 2011 autumn administration is displayed together with a fitted log-linear model, and in the lower panel a corresponding graph for the 2012 spring administration is displayed. The fitted log-linear models preserved the six first moments for both distributions, meaning that \( I = 6 \) in Equation 2. The parameterization was determined using the AIC as it has been shown to be particularly accurate for univariate distributions (Moses and Holland, 2010).

![Figure 1: The observed score distribution and log-linear model fitted to the 2011 (upper panel) and 2012 (lower panel) SweSAT data.](image)

As in the rest of the thesis, the R programming language (R Core Team, 2018) and the R package `kequate` (Andersson and Wiberg, 2017) have been used.

### 4.3.2 Calculating the score probabilities

In the second step of kernel equating, the fitted values of the presmoothing models need to be mapped from the estimated probability distributions to the vectors of estimated score probabilities \( \hat{r} \) and \( \hat{s} \). This mapping differs for log-linear and IRT models. For the former, a so-called design function (DF) is specified in accordance with the data
collection design. The DF equals

\[
\begin{pmatrix}
 r \\
 s
\end{pmatrix} = \text{DF}(P, Q)
\]

where \(P\) and \(Q\) denote the score distributions estimated by the log-linear models. The DF thus maps the data, which is often of higher dimension, into a \(J \times K\)-dimensional score probability vector.

For the EG design, \(r\) and \(s\) are directly estimated by the model defined in Equation 2. The DF is therefore a mapping from \(\Omega_J \times \Omega_K\) to itself, for which

\[
\Omega_J = \left\{ r \in \mathbb{R}^J : r_j > 0 \text{ and } \sum_j r_j = 1 \right\}
\]

and

\[
\Omega_K = \left\{ s \in \mathbb{R}^K : s_k > 0 \text{ and } \sum_k s_k = 1 \right\},
\]

so that

\[
\begin{pmatrix}
 r \\
 s
\end{pmatrix} = \text{DF}(r, s) = \begin{pmatrix}
 I_J & 0 \\
 0 & I_K
\end{pmatrix} \begin{pmatrix}
 r \\
 s
\end{pmatrix},
\]

where \(I_J\) is a \(J \times J\) identity matrix and \(I_K\) is a \(K \times K\) identity matrix.

For the NEAT and NEC design, the DF is more complex since \(P\) and \(Q\) are bivariate distributions. Specifically, the \(l\)th columns of the matrices \(P\) and \(Q\) equal

\[
\begin{pmatrix}
 p_{1l} \\
 \vdots \\
 p_{Jl}
\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}
 q_{1l} \\
 \vdots \\
 q_{Kl}
\end{pmatrix},
\]

where \(q_{kl}\) is defined analogously to \(p_{jl}\), i.e. \(q_{kl} = \text{Pr}(Y = y_k, A = a_l)\). See von Davier et al. (2004) and Wallin and Wiberg (2019) for explicit expressions of the DFs for the NEAT and NEC design.

When IRT models are used in the presmoothing step, the estimated score probabilities \(\hat{r}\) and \(\hat{s}\) are obtained in a different way. Under the assumptions following an IRT model, the data is on an item level, meaning that the total score for test-taker \(i\) equals

\[
X_i = \sum_{j^*=1}^{x_J} X_{ij^*}.
\]
The probability distribution of $X_i$ given $\theta$ can be described by a compound binomial distribution as derived by Birnbaum (1968),

$$
\Pr(X_i = x|\theta_i) = \sum \sum x_{ij} \prod_{j^*=1}^{x_{ij^*}} p_{ij^*}^{x_{ij^*}} (1 - p_{ij^*})^{1-x_{ij^*}}
$$

where $p_{ij^*}$ comes from Equation 3. The recursive algorithm described by Lord and Wingersky (1984) is most-often used for the calculation of these test score probabilities.

### 4.3.3 Continuization

When the test score probabilities $r$ and $s$ have been estimated, continuous approximations of the score CDFs are needed before the equipercentile transformation can be formed. For this purpose, let $V$ denote a continuous random variable that is independent of $X$ and such that $\mathbb{E}(V) = 0$, $\text{Var}(V) = \sigma_V^2$, and with a kernel function $K(\cdot)$ which is symmetric around zero. Further let $\mu_X = \sum_j x_j r_j$, $\sigma_X^2 = \sum_j r_j (x_j - \mu_X)^2$, and $a_X^2 = \sigma_X^2 / (\sigma_X^2 + \sigma_V^2 h_X^2)$, where $h_X > 0$ is the bandwidth. In kernel equating, the discrete score variable $X$ is replaced by the continuous random variable

$$
X(h_X) = a_X (X + h_X V) + (1 - a_X) \mu_X.
$$

The random variable $X(h_X)$ is defined such that $\mathbb{E}(X(h_X)) = \mathbb{E}(X)$ and $\text{Var}(X(h_X)) = \text{Var}(X)$. The kernel smoothing of the distribution of $X$ equals the distribution of $X(h_X)$, which in turn equals

$$
F_{h_X}(x) = \sum_j r_j K\left( \frac{x - a_X x_j - (1 - a_X) \mu_X}{a_X h_X} \right), \tag{4}
$$

where $x$ is any real value (von Davier et al., 2004). A replacement for $Y$ is done in a corresponding way to obtain $G_{h_Y}$. Returning to the empirical example using the SweSAT data, Figure 2 illustrates the continuization of the discrete score CDFs by the use of continuous functions. The kernel function $K(\cdot)$ is chosen to be $\Phi(\cdot)$, the CDF of a standard normal distribution. Other choices of kernel function, such as the uniform and logistic, have also been implemented within kernel equating (Lee and von Davier, 2011).
The bandwidth $h_X$ determines the level of smoothness of the continuous CDF approximations and can be selected in several ways, such as by the minimization of penalty functions (von Davier et al., 2004), by Silverman’s rule of thumb (Andersson and von Davier, 2014), double smoothing (Häggström and Wiberg, 2014), and likelihood cross-validation (Liang and von Davier, 2014).

Figure 2: The empirical and continuized score CDFs of the SweSAT data.

4.3.4 Equating

When the discrete score CDFs have been turned into continuous functions, denoted by $\hat{F}_{h_X}(x) = F_{h_X}(x; \hat{r})$ and $\hat{G}_{h_Y}(y) = G_{h_Y}(y; \hat{s})$, the equating function can be formed as

$$\hat{\varphi}(x) = G_{h_Y}^{-1}(F_{h_X}(x; \hat{r}); \hat{s}),$$

(5)

Under the NEAT and NEC designs, the equating function in Equation 5 can be formed using both post-stratification equating (PSE) and chained equating (CE). For PSE, the estimated distributions $F_{h_X}(x; \hat{r})$ and $G_{h_Y}(y; \hat{s})$ are calculated for the target population defined by $T = wP + (1 - w)Q$, $w \in [0, 1]$. This means that

$$r_j = w\Pr(X = x_j | P) + (1 - w)\Pr(X = x_j | Q)$$

(6)
and

\[ s_k = w \Pr(Y = y_k|P) + (1 - w) \Pr(Y = y_k|Q), \tag{7} \]

where \( P \) and \( Q \) are references to the populations for which the probabilities are calculated. Since the sample from \( P \) only have been administered test form \( X \) and the sample from population \( Q \) only have been administered test form \( Y \), the probabilities \( \Pr(X = x_j|P) \) and \( \Pr(Y = y_k|Q) \) are directly estimable from data, but \( \Pr(X = x_j|Q) \) and \( \Pr(Y = y_k|P) \) are not. However, assuming that each score distribution is population invariant given the proxy variable (for example the anchor or propensity score), they are possible to obtain by conditioning on the proxy. Specifically, the PSE assumptions are

\[
\Pr(X = x_j|A = a_l, P) = \Pr(X = x_j|A = a_l, Q) \]

and

\[
\Pr(Y = y_k|A = a_l, Q) = \Pr(Y = y_k|A = a_l, P) \]

Once the probabilities \( r_j \) and \( s_k \) have been obtained they are plugged into CDF estimators following the form of Equation 4. The equipercentile transformation of Equation 5 can thereafter be composed.

CE makes a two-stage linking from the \( X \) scores to the \( Y \) scores; first from \( X \) to \( A \) and then from \( A \) to \( Y \). Assuming that the following CDFs have been properly continuized, let \( H(\cdot) \), \( F(\cdot) \) and \( G(\cdot) \) denote the CDFs of the proxy variable of the latent ability, test score \( X \) and test score \( Y \), respectively, defined on a target population of the form \( wP + (1 - w)Q \). Let \( H_P(\cdot) \) denote the CDF of the proxy variable defined for population \( P \), let \( H_Q(\cdot) \) denote the corresponding distribution for the \( Q \) population, and let \( F_P(\cdot) \) and \( G_Q(\cdot) \) denote the corresponding CDFs for score variables \( X \) and \( Y \). To be able to express the two-stage linking of CE in the form of Equation 5, it is assumed that the links from \( X \) to the proxy, and from the proxy to \( Y \), are population invariant,

\[
H_P^{-1}(F_P(x)) = H^{-1}(F(x))
\]

and

\[
G_Q^{-1}(H_Q(a)) = G^{-1}(H(a)).
\]

It follows that an equating estimator can be composed in the form of Equation 5 as

\[
\hat{\phi}_{CE}(x) = G_{h_Y}^{-1}(F_{h_X}(x; \tilde{r}; \tilde{s})) = \hat{G}_{Q_{h_Y}}^{-1}\left( \hat{H}_{Q_{h_Y}}^{-1}\left( \hat{H}^{-1}_{F_{h_X}}(\hat{F}_{P_{h_X}}(x)) \right) \right).
\]
The estimated equating transformation for the SweSAT data is illustrated in Figure 3. In the upper panel, the definition of equivalent scores is illustrated, where for example a score point of 50 on the new test form is equivalent to a score point of 45 on the old test form. The complete equating function is illustrated in the lower panel of Figure 3.

![Diagram](image)

Figure 3: Upper panel: The definition of equivalent scores. Lower panel: The estimated equating function for the SweSAT data.

### 4.3.5 Evaluation of the equating function

As with any statistical estimator, the kernel equating estimator is subject to sampling variability. One of the most common ways of evaluating the equating function is by calculating the SEE. Using that the kernel equating estimator is continuous and differentiable with respect to the score probabilities, the asymptotic standard error has been derived using the delta method. The SEE of \( \varphi(x) \) following this approach equals

\[
\text{SEE}_Y^\Delta(x) = ||J_\varphi J_{DF} C||,
\]

where \(|| \cdot ||\) denotes the Euclidean norm, \(J_\varphi\) equals the Jacobian vector of \(\varphi\), \(J_{DF}\) denotes the Jacobian matrix of the design function and \(C\) is a matrix following of a Cholesky decomposition of the covariance matrix of the score probabilites (von Davier et al., 2004).
Liou and Cheng (1995) and Liou et al. (1997) used a different approach and derived the asymptotic SEE using the Bahadur representation of sample quantiles (Bahadur, 1966). Following the formulation of Ghosh (1971), the Bahadur representation states that for a continuous random variable $Y$ with CDF $G(y)$, quantile function $\xi_p$, $0 < p < 1$, and empirical CDF $G_N(y)$, the empirical quantile function $\hat{\xi}_p$ can be represented as

$$\hat{\xi}_p = \xi_p + \frac{p - G_N(\xi)}{G'(\xi)} + o_p\left(N^{-1/2}\right),$$

(9)
given that $G$ is once differentiable at $\xi = G^{-1}(p)$, where $N$ denotes the sample size and $o_p$ denotes convergence in probability.

Liou and Cheng (1995) established a Bahadur representation of the traditional equipercentile transformation, and later Liou et al. (1997) did so under the kernel equating framework. In the latter case, a general formula for the SEE of the kernel equating estimator that holds for all common data collection designs except for the CE approach of the NEAT design was derived:

$$\text{SEE}_B^Y(x) = \frac{1}{G'_h} \left\{ \text{Var}(F_{h_x}) + \text{Var}(G_{h_Y}) - 2\text{Cov}(F_{h_x}, G_{h_Y}) \right\}^{1/2},$$

(10)

For CE, the SEE using the Bahadur representation equals

$$\text{SEE}_Y^B(x) = \left\{ \left( \frac{1}{H'_P(\varphi_A(x))} \right)^2 \left[ \frac{H'_Q(\varphi_A(x))}{G'_Q(\varphi_{CE}(x))} \right]^2 \text{Var}[F_P(x) - H_P(\varphi_A(x))] + \left( \frac{1}{G'_Q(\varphi_{CE}(x))} \right)^2 \text{Var}[H_Q(\varphi_A(x)) - G_Q(\varphi_A(x))] \right\}^{1/2},$$

(11)

where $\varphi_A(x)$ denotes the equating function mapping $X$ to $A$.

When the variance and covariance terms in Equations 10 and 11 are estimated using Taylor expansions, i.e., the delta method, it can be shown that the SEE in Equation 8 is, for all data collection designs, equivalent to the SEE resulting from using the Bahadur representation.

The SEE using the SweSAT data is illustrated in Figure 4. It is evident that the SEE gives a measure of uncertainty at each equated score. It also exhibits the typical peaks at the lower and higher ends of the score scale since there are only a few test-takers with those scores.
Another common evaluation measure is the percent relative error (PRE; von Davier et al. 2004), which compares the \( p \)th moment of the equated scores and the \( Y \) scores. The PRE is defined as

\[
\text{PRE}(p) = 100 \left( \frac{\hat{\mu}_Y - \mu_p(Y)}{\mu_p(Y)} \right),
\]

where \( \mu_p(Y) = \sum_k (y_k)^p s_k \) and \( \hat{\mu}_Y = \sum_j (\hat{\phi}(x_j))^p r_j \).

5 Summary of papers

5.1 Paper I

In Paper I, the presmoothing of bivariate test score distributions is studied. For this purpose, there are two choices to make: what class of model to use and what model fit index to use to select the functional form of the model. Within kernel equating, two ways of presmoothing has been suggested, one using log-linear models and one using IRT models. We study the most common model fit indices, namely the AIC, BIC and likelihood ratio chi-square statistic and investigate whether these indices result in different model parameterizations. If so, it is investigated.
how the resulting presmoothing models of the different model selection criteria introduce differences in terms of equated scores. Using both real and simulated data, it is shown that different model selection criteria can result in different model parameterizations, that in turn result in equated scores with differences of practical importance. From the simulation study, the BIC shows the best possibility to select the correct model for bivariate distributions.

5.2 Paper II

In Paper II, the continuization of the estimated score CDFs is studied. Using kernel functions to approximate the discrete distributions with continuous functions, a bandwidth that determines the smoothness needs to be selected. There are currently four bandwidth selection methods proposed within kernel equating, and all of them are evaluated using both real and simulated data. Two novel bandwidth selection methods that use leave-one-out cross-validation are also introduced. The results show that the different methods yield clearly different bandwidths, but that the influence on the equated scores is marginal. Therefore, it does not appear to be a strong indication that the choice of bandwidth is of critical importance within kernel equating.

5.3 Paper III

In Paper III the NEC design is studied, for which background information on the test-takers is used to adjust for ability differences between the test groups to be equated. The covariates are gathered in a propensity score to solve the dimensionality problem that easily occurs if more than a few covariates are included. The test scores are thereafter equated with respect to the propensity score. Both a PSE and CE estimator that use propensity scores are derived, together with their respective SEE. The estimators are evaluated in a simulation study where it is shown that both estimators perform better than if an EG design would have been implemented. The estimators moreover perform equally well or better in terms of bias and root mean squared error compared to if anchor scores would have been available and used.
5.4 Paper IV

In paper IV, the SEE is studied. The SEE for kernel equating has normally been calculated using the delta method. In Paper IV, the Bahadur representation of sample quantiles for deriving the SEE is revisited. Previous results of the Bahadur representation for the SEE are extended to include all common data collection designs, and for a general class of kernel functions. Most importantly it is shown that the Bahadur SEE and the delta method SEE are equivalent for all data collection designs when the unknown components of the Bahadur SEE are estimated using Taylor expansions. Another way of estimating the components are included in the paper and is studied in an empirical example where comparisons to the delta method SEE are made. The overall results show only small differences between the methods.

6 Final remarks and further research

This thesis extends the kernel method of test score equating by studying presmoothing, continuization, ability adjustment using covariates and SEE calculation.

Since the IRT models have recently been implemented within kernel equating, it is of interest to further study the implications of the five steps of kernel equating using this class of models. There are other presmoothing models suggested within traditional equipercentile equating so there is a potential to investigate the implementation of such models within the kernel equating framework.

In Paper II, the selection of bandwidth for the continuization step was investigated. Out of the bandwidth selection methods that were studied, only the Silverman’s rule of thumb method have SEEs that are adjusted for the uncertainty in the bandwidth selection. Future research therefore should derive SEE formulas that take such uncertainty into account also for the other bandwidth selection methods.

In Paper III, logistic regression was used to estimate the propensity score for each test-taker. It would be of interest to study other estimators like probit regression and non-parameteric estimators. It would furthermore be interesting to study the effect of model misspecification on the equated scores. The estimation of propensity scores is also a source of uncertainty, and SEE formulas that take this into account is
another possible study for the future.

In Paper IV, the Bahadur representation of sample quantiles was studied for the derivation of the SEE. It results in a formula that consists of two variance terms and one covariance term. A future study of interest would be to use this fact in the modeling phase and for example use the variance sizes as sources of information about how to optimally select for example the kernel function and bandwidth.

References


