UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 106

Contributions to Kernel Equating

BJÖRN ANDERSSON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2014

ISSN 1652-9030
ISBN 978-91-554-9089-8
urn:nbn:se:uu:diva-234618

Dissertation presented at Uppsala University to be publicly examined in Sal IV, Universitetshuset, Biskopsgatan 3, Uppsala, Friday, 12 December 2014 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Jorge González (Pontificia Universidad Católica de Chile).

Abstract

Andersson, B. 2014. Contributions to Kernel Equating. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 106. 24 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9089-8.

The statistical practice of equating is needed when scores on different versions of the same standardized test are to be compared. This thesis constitutes four contributions to the observed-score equating framework kernel equating.

Paper I introduces the open source R package kequate which enables the equating of observed scores using the kernel method of test equating in all common equating designs. The package is designed for ease of use and integrates well with other packages. The equating methods non-equivalent groups with covariates and item response theory observed-score kernel equating are currently not available in any other software package.

In paper II an alternative bandwidth selection method for the kernel method of test equating is proposed. The new method is designed for usage with non-smooth data such as when using the observed data directly, without pre-smoothing. In previously used bandwidth selection methods, the variability from the bandwidth selection was disregarded when calculating the asymptotic standard errors. Here, the bandwidth selection is accounted for and updated asymptotic standard error derivations are provided.

Item response theory observed-score kernel equating for the non-equivalent groups with anchor test design is introduced in paper III. Multivariate observed-score kernel equating functions are defined and their asymptotic covariance matrices are derived. An empirical example in the form of a standardized achievement test is used and the item response theory methods are compared to previously used log-linear methods.

In paper IV, Wald tests for equating differences in item response theory observed-score kernel equating are conducted using the results from paper III. Simulations are performed to evaluate the empirical significance level and power under different settings, showing that the Wald test is more powerful than the Hommel multiple hypothesis testing method. Data from a psychometric licensure test and a standardized achievement test are used to exemplify the hypothesis testing procedure. The results show that using the Wald test can provide different conclusions to using the Hommel procedure.

Keywords: observed-score test equating, item response theory, R, equipercentile equating, asymptotic standard errors, non-equivalent groups with anchor test design

Björn Andersson, Department of Statistics, Uppsala University, SE-75120 Uppsala, Sweden.

ISSN 1652-9030
ISBN 978-91-554-9089-8
urn:nbn:se:uu:diva-234618 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-234618)

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Andersson, B., Bränberg, K. and Wiberg, M. (2013). Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software*, **55(6)**, 1-25.

II Andersson, B. and von Davier, A. A. (2014). Improving the Bandwidth Selection in Kernel Equating. *Journal of Educational Measurement*, **51(3)**, 223-238.

III Andersson, B. and Wiberg, M. (2014). Item Response Theory Observed-Score Kernel Equating. *Submitted*.

IV Andersson, B. (2014). An Evaluation of Hypothesis Testing Methods for Equating Differences in Kernel Equating. *Manuscript*.

Contents

1 Introduction 7
- 1.1 Data and data collection designs 7
- 1.2 Test equating 8
- 1.3 Item response theory observed-score equating 10
- 1.4 The kernel equating framework 10
- 1.5 Bandwidth selection in kernel equating 13
- 1.6 Choosing between different equating functions 14

2 Objective of the thesis 16

3 Summary of papers 17
- 3.1 Performing the Kernel Method of Test Equating with the Package kequate 17
- 3.2 Improving the Bandwidth Selection in Kernel Equating 17
- 3.3 Item Response Theory Observed-Score Kernel Equating 18
- 3.4 An Evaluation of Hypothesis Testing Methods for Equating Differences in Kernel Equating 19

4 Conclusions 21

Acknowledgements 22

References 23

1. Introduction

Test equating is the statistical procedure by which scores on two separate tests on the same topic are related. The major area of application for test equating is in standardized testing, where a certain ability or abilities are measured by a test designed for that specific purpose. Standardized testing is used in many different settings, such as in evaluating the performance of a particular population or when evaluating the performance of an individual. Examples of the former case are PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) and examples of the latter case are the Swedish Scholastic Aptitude Test, SAT and TOEFL (Test of English as a Foreign Language). For these tests, different versions of the same test are administered at different points in time. Often the different versions are administered to groups of people from populations which differ from one instance to the next. For tests which are meant to evaluate the performance of an individual, it is necessary to relate the scores on different versions of the same test in order to compare the scores of the individuals taking the different versions. Test scores are the basis for admission to university programs and are used in deciding between passing or failing a certification. Hence it is of high importance that the scores are comparable from different versions to ensure that the test-takers are evaluated fairly. In designing a test, the overall item difficulty is meant to be equal across the different test versions in order to facilitate comparisons between individuals taking the different versions. In practice, two tests consisting of different items are rarely perfectly equal in difficulty. This leads to the purpose of test equating: to ensure that scores from different administrations of the same test are comparable.

1.1 Data and data collection designs

Standardized tests often consist of multiple choice items which are scored as either true (denoted as 1) or false (denoted as 0). Such items are said to be dichotomous. The items can also be scored in multiple categories and such items are called polytomous items. The score on a polytomous item with m possible categories is denoted $0, 1, \ldots, m-1$. The data resulting from a test administration thus consist of a sequence of numbers denoting the score on each item for each individual. Often, the data used in equating are the summed scores over all the items for each individual.

In equating, there are many different data collection designs used to relate scores on two versions of the same standardized test. Let X and Y denote the two different versions of the same standardized test and let P and Q denote the populations, possibly identical, which the groups taking the respective tests are from. The descriptions of the data collection designs given here are adapted from Kolen and Brennan (2014) and von Davier et al. (2004).

In the equivalent groups (EG) design, different individuals from a common population take each of the versions X and Y at different points in time, enabling the direct comparison of the scores on the different versions. In a single group (SG) design, the same individuals take both tests X and Y, also enabling the direct comparison of the scores on the two versions. This design is however afflicted by a possible practice effect since one test is taken before the other. A way to mitigate this effect is to have half of the group take test X first and have the other half take test Y first. Such a design is called a counterbalanced (CB) design.

In the non-equivalent groups design, individuals from different populations take the versions X and Y at different points in time. In this design, it is not possible to directly relate the scores on versions X and Y since the groups do not come from the same population and may differ in ability. In order to relate the scores of X and Y, common items can be administered to each group in addition to X and Y. These common items constitute the anchor test A, which may be part of the main tests X and Y (internal anchor) or may be given separately (external anchor). The information that the common items provide is used to relate the scores on X and Y. The described design is called a non-equivalent groups with anchor test (NEAT) design. If an anchor test is not administered, it is still possible to equate two versions X and Y if covariates which are correlated with the test scores are available for the individuals. The design in this setting is called the non-equivalent groups with covariates (NEC) design (Bränberg and Wiberg, 2011).

1.2 Test equating

The object of an equating transformation is to relate the scores on two different versions of the same test. Such a transformation is a function of the observed data and estimated statistical models. Thus, equating is a statistical procedure by which scores on different test forms are adjusted so that the scores from these test forms can be used interchangeably (Kolen and Brennan, 2014).

In order for a transformation to be called an equating function, five requirements have been identified (Lord, 1980; von Davier et al., 2004; Kolen and Brennan, 2014). First, the equal construct requirement says that the tests to be equated should measure the same underlying construct. Second, the equal reliability requirement means that the tests to be equated should have equal reliability. Third, the symmetry requirement states that the equating transfor-

mation should be symmetrical, i.e. for tests X and Y there should not be a difference in equating X to Y compared to equating Y to X. Fourth, the equity requirement means that it should not matter to an individual whether test X or test Y is administered. Fifth, the population invariance requirement states that the equating transformation should be identical no matter which population were administered the tests X and Y. In practice, it is not possible to guarantee that the requirements of equal reliability, equity and population invariance hold but the tests are meant to be designed such that these requirements are fulfilled. There are ways to empirically assess that the requirements are satisfied for a given test administration, see e.g. Dorans and Holland (2000).

There are many different procedures that can be used to conduct an equating of two test forms. These procedures can in large part be separated into two approaches: observed-score equating and true score equating. The observed score is the score which a given individual receives after taking the test. The true score equals the observed score plus a random, unobserved, error term. When equating true scores, the object is to find the transformation applied to the true score on test X such that the expected value of the transformation equals the expected value of the true score on test Y. In observed-score equating, the object is to find the test Y equivalent observed score of test X. In this thesis the focus is only on observed-score equating.

Observed-score equating is itself divided into two main approaches. The first is called linear observed-score equating, which means that there is a linear relationship between the observed scores on test X and test Y. Let X be a test with k_X dichotomous items. The possible score values on test X are then $\{x_1, \ldots, x_{k_X+1}\}$. The linear equating function is defined as

$$e_{Y(\mathrm{LIN})}(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}(x - \mu_X), \tag{1.1}$$

where μ_X and μ_Y are the means and σ_X and σ_Y are the standard deviations of the test score distributions for tests X and Y, respectively. The second type of observed-score equating function is called the equipercentile equating function, defined as

$$e_Y(x) = G^{-1}[F(x)], \tag{1.2}$$

where $F(\cdot)$ and $G(\cdot)$ are the cumulative distribution functions for tests X and Y, respectively. However, in standardized testing, scores are usually integer-valued and hence the distribution functions in Equation 1.2 are not continuous. For this reason continuous approximations must be defined. Before the advent of kernel equating, these continuous approximations were calculated using linear interpolation (Angoff, 1984). In linear interpolation, the continuous

approximation to $F(\cdot)$ is defined as

$$F_{\text{LI}}(x;\boldsymbol{\alpha}_X) = \begin{cases} 0 & \text{if } x \leq x_1 - 0.5 \\ \sum_{j=1}^{k_X+1} r_j + [x - (x_k - 0.5)]r_k & \text{if } x_1 - 0.5 < x \leq x_{k_X+1} + 0.5 \\ 1 & \text{if } x > x_{k_X+1} + 0.5 \end{cases}, \tag{1.3}$$

where k equals the nearest integer to x, r_j is the score probability for score point j and r_k is the score probability for score point k. Two drawbacks to using linear interpolation are that the resulting distribution function is not everywhere differentiable and that the variance of the original distribution is not preserved (see e.g. von Davier et al. (2004)).

1.3 Item response theory observed-score equating

Item response theory (IRT) is a commonly used statistical method to model the responses to the items of a standardized test. In unidimensional IRT the probabilities to answer each item on a test correctly are assumed to be functions of an underlying latent variable θ and item parameters which determine the shape of the functions (Hambleton and Swaminathan, 1985). A popular IRT model is the three-parameter logistic model (Lord, 1980), where the probability to answer the dichotomous item l on the test X correctly is modelled as

$$P_{Xl}(\theta) = c_{Xl} + \frac{1 - c_{Xl}}{1 + \exp\left[-a_{Xl}(\theta - b_{Xl})\right]}, \tag{1.4}$$

where the parameter a_{Xl} denotes the discrimination of the item (how well the item separates between low ability and high ability individuals), the parameter b_{Xl} the difficulty of the item and c_{Xl} is the guessing parameter for the item (the lower bound for the probability of answering the item correctly). Setting $c_{Xl} = 0$ retrieves the two-parameter logistic model. The item parameters from an IRT model can be used to calculate the score probabilities for each summed score on the test. These score probabilities can then be used to conduct an observed-score equating (Lord and Wingersky, 1984). Hence, the IRT model estimation can be viewed as a pre-smoothing step in the equating process.

1.4 The kernel equating framework

Kernel equating was first introduced by researchers at Educational Testing Service in the late 1980s where it was described for the EG, SG and NEAT post-stratification equating (NEAT PSE) designs (Holland et al., 1989; Holland and Thayer, 1989). At the time, kernel equating was unique in providing standard errors of equating when using pre-smoothing with log-linear models.

The kernel method of test equating was further developed in the early 2000s, again at Educational Testing Service, when kernel equating was extended to include the CB and NEAT chain equating (NEAT CE) designs. The concept of standard error of equating difference (SEED) was also introduced. This research was summarized in the book *The Kernel Method of Test Equating* (von Davier et al., 2004) and in von Davier (2013).

The kernel method of test equating has typically been described as a procedure comprising five steps:

1) Pre-smoothing of the score probabilities

In order to reduce the variance and get a more stable equating function a parametric model is most often fitted to the observed data. This procedure is called pre-smoothing. The original proposal in kernel equating was to calculate the summed score for each individual and fit log-linear models to smooth out the resulting score probabilities (Holland and Thayer, 1989, 2000). Another option is to fit an IRT model from the responses for each individual on each item and from this model calculate the implied, smoothed, score probabilities. It is possible to avoid the pre-smoothing step and use the observed data directly. However, pre-smoothing has been shown to be effective in improving the accuracy of the resulting equating (Kolen and Brennan, 2014).

2) Calculation of the score probabilities

After fitting the parametric model to the data, the resulting parameter estimates are used to calculate the score probabilities required for each design. This step differs somewhat depending on the pre-smoothing method used. For the EG and NEAT CE designs, the step is identical for pre-smoothing using either log-linear or IRT models. In either pre-smoothing method, the marginal score probability vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ for tests X and Y respectively are calculated for the EG design and the marginal score probability vectors $\boldsymbol{r}_P, \boldsymbol{t}_P, \boldsymbol{s}_Q$ and $\boldsymbol{t}_Q$ for tests X and A on P and tests Y and A on Q, respectively, are calculated for the NEAT CE design. The calculation of the score probabilities $\boldsymbol{r}$ and $\boldsymbol{s}$ in the SG design and the score probabilities $\boldsymbol{r}_S$ and $\boldsymbol{s}_S$ in the NEAT PSE design differs between the log-linear and IRT methods. For the log-linear method, estimated bivariate distributions are used to calculate the required score probabilities whereas for the IRT method concurrent calibration (SG) or equating coefficients (NEAT PSE) are used. The NEAT PSE case is a bit different than the other methods, since the resulting score probabilities $\boldsymbol{r}_S$ and $\boldsymbol{s}_S$ are defined for a synthetic population S, a mixture of the two populations P and Q:

$$S = w_S \times P + (1 - w_S) \times Q, \tag{1.5}$$

where $w_S \in [0, 1]$.

3) Calculation of the continuous approximation to the discrete test score distribution

When the score probabilities have been calculated, the resulting discrete distribution functions must be converted to continuous distribution functions in order to conduct the equating. This step is identical for all methods of presmoothing and for using the observed data directly. Consider an EG design, where the discrete distribution function for test X with k_X dichotomous items is $F(x;\boldsymbol{r})$. The kernel method continuous approximation to $F(x;\boldsymbol{r})$ is

$$F_{h_X}(x;\boldsymbol{r}) = \sum_{j=1}^{k_X+1} r_j \Phi\left(\frac{x - a_X x_j - (1-a_X)\mu_X}{a_X h_X}\right), \tag{1.6}$$

where r_j is the score probability for the j-th score value, $\Phi(\cdot)$ denotes the standard normal distribution function, x_j is the j-th score value, μ_X is the mean of the test scores, h_X is the bandwidth and

$$a_X = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + h_X^2}}\,, \tag{1.7}$$

where σ_X^2 is the variance of the test scores. The bandwidth h_X is discussed in Section 1.5. Let $G_{h_Y}(\cdot;\boldsymbol{s})$ denote the continuous approximation for test Y. In the NEAT CE design, let $F_{Ph_X}(\cdot;\boldsymbol{r}_P)$ and $H_{Ph_{AP}}(\cdot;\boldsymbol{t}_P)$ be the continuous approximations of the distribution functions for tests X and A on P and let $G_{Qh_Y}(\cdot;\boldsymbol{s}_Q)$ and $H_{Qh_{AQ}}(\cdot;\boldsymbol{t}_Q)$ be the continuous approximations of the distribution functions for tests Y and A on Q. For the NEAT PSE design, define $F_{Sh_X}(\cdot;\boldsymbol{r}_S)$ and $G_{Sh_Y}(\cdot;\boldsymbol{s}_S)$ as the continuous approximations for tests X and Y on the synthetic population S. Each continuous approximation is calculated in the same way as that in Equation 1.6. The kernel equating framework enables the usage of other kernels than the Gaussian kernel used in Equation 1.6, such as the logistic and uniform kernels (Lee and von Davier, 2011). The continuous approximations provided by the kernel method are differentiable and preserve both the mean and the variance of the original test score distributions (von Davier et al., 2004).

4) Equating

At each score point of test X an equated value is calculated according to the specific design. The function corresponding to this transformation is called the equating function. In the EG design, the equating function from X to Y is the inverse of the continuous approximation $G_{h_Y}(\cdot;\boldsymbol{s}_Q)$ evaluated at the value of the continuous approximation for test X evaluated at the score point x:

$$e_{Y(\mathrm{EG})}(x;\boldsymbol{r},\boldsymbol{s}) = G_{h_Y}^{-1}\left[F_{h_X}(x;\boldsymbol{r});\boldsymbol{s}\right]. \tag{1.8}$$

In the NEAT CE design, the equating function is a composite function of the four continuous approximations for each test and population combination:

$$e_{Y(\mathrm{CE})}(x;\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q) = G^{-1}_{Qh_Y}\left(H_{Qh_{AQ}}\left(H^{-1}_{Ph_{AP}}\left(F_{Ph_X}\left(x;\boldsymbol{r}_P\right);\boldsymbol{t}_P\right);\boldsymbol{t}_Q\right);\boldsymbol{s}_Q\right). \tag{1.9}$$

For the NEAT PSE design, the equating is conducted with respect to the synthetic population S. The equating function in this design is defined as:

$$e_{Y(\mathrm{PSE})}(x;\boldsymbol{r}_S,\boldsymbol{s}_S) = G^{-1}_{Sh_Y}\left[F_{Sh_X}(x;\boldsymbol{r}_S);\boldsymbol{s}_S\right]. \tag{1.10}$$

It is useful to consider the equating function as a vector function for all score points $x_1,\ldots,x_{k_X+1}$ on test X, defined for any equating design and any method of pre-smoothing. Hence, denote the general multivariate equating function for a specific design D as

$$\boldsymbol{e}_{Y(\mathrm{D})}(\mathbf{x};\boldsymbol{\tau}) = \left[e_{Y(\mathrm{D})}\left(x_1;\boldsymbol{\tau}\right) \quad \ldots \quad e_{Y(\mathrm{D})}\left(x_{k_X+1};\boldsymbol{\tau}\right)\right]', \tag{1.11}$$

where $\boldsymbol{\tau}$ is the vector of parameters in the pre-smoothing model.

5) Calculating the standard error of equating

In practice, the equating function is unknown and must be estimated. Denote the estimator of the general multivariate equating function $\hat{\boldsymbol{e}}_{Y(\mathrm{D})}(\mathbf{x};\hat{\boldsymbol{\tau}})$. The estimator is subject to sampling variability and hence calculating the variance of the estimator is desirable. Let n be the sample size. Under an assumption of asymptotic normality of the estimator of the score probabilities, large sample approximations using the delta method are used to calculate the variance of $\hat{\boldsymbol{e}}_{Y(\mathrm{D})}(\mathbf{x};\hat{\boldsymbol{\tau}})$ (Ferguson, 1996). The delta method can be used since $\hat{\boldsymbol{e}}_{Y(\mathrm{D})}(\mathbf{x};\hat{\boldsymbol{\tau}})$ is continuous and differentiable with respect to the score probabilities. Thus, as $n \to \infty$,

$$\sqrt{n}\left(\hat{\boldsymbol{e}}_{Y(\mathrm{D})}(\mathbf{x};\hat{\boldsymbol{\tau}}) - \boldsymbol{e}_{Y(\mathrm{D})}(\mathbf{x};\boldsymbol{\tau})\right) \to N\left(0,\boldsymbol{\Sigma}_{\hat{\boldsymbol{e}}_{Y(\mathrm{D})}(\mathbf{x};\hat{\boldsymbol{\tau}})}\right). \tag{1.12}$$

Formulas which can be used to calculate $\boldsymbol{\Sigma}_{\hat{\boldsymbol{e}}_{Y(\mathrm{D})}(\mathbf{x};\hat{\boldsymbol{\tau}})}$ when using pre-smoothing with log-linear models in the EG, SG, CB, NEAT CE and NEAT PSE designs are given in von Davier et al. (2004).

1.5 Bandwidth selection in kernel equating

The kernel method of test equating requires the selection of bandwidth parameters which determine the features of the resulting continuous approximations to the discrete test score distributions. A small bandwidth puts more emphasis on the value at which the function is evaluated whereas a larger bandwidth

is influenced to a higher degree by the adjacent score points. A higher bandwidth thus produces smoother distribution functions. If the bandwidths are set to very large numbers, in von Davier et al. (2004) defined as 10 times the standard deviations of the test scores, the resulting equating function will closely match the linear equating function. Although the bandwidth can be set beforehand by the practitioner to any desired value, in kernel equating two main data-driven methods of selecting the bandwidth have been proposed. Both methods utilize penalty functions to select a bandwidth which is in some sense optimal for a given input of score probabilities. Let $\hat{r}_j$ denote the estimated score proportion for score value $j \in \{1, \ldots, k_X + 1\}$ and let $\hat{F}'_{h_X}(\cdot)$ and $\hat{F}''_{h_X}(\cdot)$ denote the first and second derivatives of the estimated continuous distribution function, respectively. The first method selects the bandwidth by minimizing the function

$$\text{PEN}_1(h_X) = \sum_{j=1}^{k_X+1} \left[\hat{r}_j - \hat{F}'_{h_X}(x_j)\right]^2, \tag{1.13}$$

which gives a density function that closely resembles the estimated or observed proportions. The second method selects the bandwidth by minimizing the function

$$\text{PEN}(h_X) = \text{PEN}_1(h_X) + \kappa \sum_{j=1}^{k_X+1} A_j \,, \tag{1.14}$$

where κ is a constant usually set to 1 and $A_j = 1$ if $\hat{F}''_{h_X}(x_j - \omega) > 0$ and $\hat{F}''_{h_X}(x_j + \omega) < 0$ or if $\hat{F}''_{h_X}(x_j - \omega) < 0$ and $\hat{F}''_{h_X}(x_j + \omega) > 0$, where ω is a constant typically set to $\omega = 1/4$. The second method penalizes for irregularities around each score point, providing a more smooth density function. Although the bandwidths are influenced by the features of the estimated score probabilities and selection of the bandwidths will vary for each data set, the bandwidth selection was not taken into account in the formulas for the standard errors of equating which were provided in von Davier et al. (2004).

1.6 Choosing between different equating functions

For a given data set in a particular design the pre-smoothing method and the type of equating have to be decided. To guide the selection of the particular equating method, it is possible to look at the model fit of the pre-smoothing models considered, to compare the standard errors between the different equating methods and to consider various equating criteria (Kolen and Brennan, 2014).

Within the kernel equating framework, it has been suggested to look at the Percent Relative Error (PRE) to help decide which equating function should

be used. The PRE for the p-th moment for the equated distribution, X to Y, is defined as

$$\text{PRE}(p) = 100\frac{\mu_p\left[e_Y(\mathbf{X})\right] - \mu_p(\mathbf{Y})}{\mu_p(\mathbf{Y})}\,, \tag{1.15}$$

where $\mu_p(\mathbf{Y}) = \sum_k (y_k)^p s_k$ and $\mu_p\left[e_Y(\mathbf{X})\right] = \sum_j \left[e_Y(x_j)\right]^p r_j$, where s_k and r_j are the estimated or observed proportions corresponding to each score value y_k or equated value $e_Y(x_j)$, respectively (von Davier et al., 2004). A PRE closer to zero matches the p-th moment between the observed distribution and the equated distribution better, which is desirable.

Additionally, it is possible to consider the SEED between two equating functions $\hat{e}_{\text{Y1}}(x)$ and $\hat{e}_{\text{Y2}}(x)$ which are derived from the same pre-smoothing model, defined as

$$\text{SEED}_{\hat{e}_{\text{Y1}-\text{Y2}}}(x) = \sqrt{\text{Var}\left(\hat{e}_{\text{Y1}}(x) - \hat{e}_{\text{Y2}}(x)\right)}\,. \tag{1.16}$$

For instance, equatings in a NEAT design with log-linear pre-smoothing using CE and PSE can be compared. The SEED has also been generalized to the multivariate case and hypothesis tests can be conducted for the equating differences of two equating functions (Rijmen et al., 2011).

2. Objective of the thesis

The objective of the thesis has been to extend the kernel equating framework in multiple ways. Firstly by implementing an open source software package for kernel equating, which can be used freely by practitioners and researchers. Additionally, the bandwidth selection method has been improved and a data-driven bandwidth choice has been introduced which enables the standard errors of equating to incorporate the variability in the bandwidth selection. Furthermore, IRT observed-score equating has been incorporated in the kernel equating framework. Lastly, the equating function has been generalized to the multivariate case and hypothesis testing between different equating methods for two separate pre-smoothing settings has been investigated.

3. Summary of papers

3.1 Performing the Kernel Method of Test Equating with the Package **kequate**

In recent years, it has become increasingly popular to use open source software such as R for statistical analysis. The R package **kequate** which implements the kernel method of test equating is presented in paper I. The package is released under the GPL-3 license and can be downloaded at `http://cran.r-project.org/package=kequate`. While the kernel method of test equating has been implemented in the proprietary software package LOGLIN/KE (Chen et al., 2011) and the C library Equating Recipes (Brennan et al., 2009) there has not previously existed an accessible open source software package to conduct kernel equating.

The implementation of the kernel method in **kequate** enables observed-score equating using the EG, SG, CB, NEAT and NEC designs. Both data which have been smoothed using log-linear models and unsmoothed data are supported. Additionally, IRT observed-score kernel equating is included. For data smoothed by log-linear models, **kequate** provides a convenient way of using objects created by the R function `glm()` (**stats** R Development Core Team, 2013). For IRT observed-score kernel equating, support is provided for IRT model estimation with the package **ltm** (Rizopoulos, 2006). There also exists an option to select the type of kernel to use in the continuous approximation step, with Gaussian, logistic and uniform kernels supported.

The package offers various ways to customize the analysis by selecting the bandwidth parameters manually and specifying the parameters used in the different kernels. Using **kequate** it is also easy to compare equatings with the built-in functions to calculate the SEED and to plot the results. In the paper, kernel equating is illustrated by equating tests in the EG, NEAT and NEC designs. The NEC design and IRT observed-score kernel equating are currently not available in any other software package. Due to the relative ease of extending the kernel method of test equating, new additions to this framework are expected to be implemented in the package in the future.

3.2 Improving the Bandwidth Selection in Kernel Equating

Paper II of the thesis discusses the most commonly used bandwidth selection methods currently used in kernel equating and proposes a new bandwidth se-

lection method for Gaussian kernels based on what is known as Silverman's rule of thumb. Using a variant of Silverman's rule of thumb (Silverman, 1986), the bandwidth proposed for usage in kernel equating equals a function of the standard deviation of the test scores, σ_X, and the sample size N_X,

$$h = \frac{9\sigma_X}{\sqrt{100N_X^{2/5} - 81}}. \tag{3.1}$$

This way of selecting the bandwidth parameters provides sufficient smoothing for erratic test score distributions while providing a way to account for the variability in the bandwidth selection method in the standard error derivations. Unlike previous methods, using the above variant of Silverman's rule of thumb provides analytical standard errors which do not underestimate the true standard errors. The updated standard error derivations are given in the paper.

Using the formula in Equation 3.1 to find the bandwidth generally results in bandwidths which are slightly larger than when employing the bandwidth selection method using penalty functions. A larger bandwidth implies an increase in the bias in the kernel method. However, as shown in the paper, in equating this bias does not manifest itself greatly compared to the commonly used methods. When compared to the full penalty function which has typically been used with non-smooth data, the method based on Silverman's rule of thumb given in Equation 3.1 provides similar bandwidths to the full penalty function. Selecting the bandwidth parameters by using the full penalty function is shown through a bootstrap analysis to have a large effect on the standard error of equating at the extreme values. Overall, the proposed method provides similar equating functions to the previous methods while being less computationally intensive and having analytical standard errors of equating which are not underestimated.

3.3 Item Response Theory Observed-Score Kernel Equating

When kernel equating was first introduced, pre-smoothing of the score probabilities was conducted using log-linear models. However, any method of pre-smoothing can be utilized in the kernel equating framework with the asymptotic results intact provided that the estimator of the score probabilities is asymptotically normally distributed. In Ogasawara (2003) IRT observed-score equating using traditional equipercentile equating was introduced. Building on the results of Ogasawara (2003), observed-score kernel equating with the two-parameter logistic and three-parameter logistic IRT models is introduced in Paper III. IRT observed-score kernel equating in the NEAT CE and NEAT PSE designs is presented and the asymptotic covariance matrices for the equating functions are derived for each design. The results generalize the work of

Ogasawara (2003) by considering a vector-valued equating function and by allowing for an arbitrary kernel in estimating the continuous approximations to the discrete distribution functions. It is also shown that the asymptotic results apply for the recently proposed IRT local kernel equating method (Wiberg et al., 2014).

The provided derivations are verified with simulations for the two-parameter and three-parameter logistic models in the NEAT CE and NEAT PSE designs. With the NEAT PSE design, both moment methods and test characteristic curve methods for estimating the equating coefficients are considered. The results show that the asymptotic standard errors are accurate for sample sizes as low as 500 when using the two-parameter logistic model with CE and with PSE using the test characteristic curve methods. The three-parameter logistic model works well only with sample sizes as large as 3000. Compared to the two-parameter logistic model, the standard errors of equating are about 25-35% larger for the three-parameter logistic model with sample size 3000. Data from a standardized achievement test are used to illustrate the methods in a practical setting. A comparison to equating with log-linear models is included, showing that the IRT methods offer lower standard errors of equating for lower and higher score points.

3.4 An Evaluation of Hypothesis Testing Methods for Equating Differences in Kernel Equating

In paper IV, the asymptotic results in paper III are used to conduct hypothesis tests of equating differences for IRT observed-score kernel equating using Wald tests (Wald, 1943). These hypothesis tests can be conducted across more score points than previously described methods using log-linear models (Rijmen et al., 2011) since the covariance matrix of the equating difference has full rank when using IRT models. In addition to introducing hypothesis tests using IRT models, simulations are conducted to evaluate the hypothesis testing of equating differences when using log-linear models, which has previously not been done. The tests are evaluated in a NEAT design by conducting simulations under the null and alternative hypotheses and recording the rejection rates for hypotheses of equality of the NEAT CE and NEAT PSE equating functions for different score ranges. The Wald test is compared to the Hommel method (Hommel, 1988), which is an alternative multiple hypothesis testing procedure that works better in the given setting than e.g. the Bonferroni correction.

The results show that a large sample size is required in order to attain the correct significance level when simulating under the null hypothesis. Across eight score points the empirical significance level did not attain the nominal level even with a very large sample size. For sample sizes which are interesting for practical use, the test is undersized. Alternative hypotheses corresponding to different degrees of violations to the null hypothesis are considered showing

that the power of the tests is good for sample sizes 3000 and up. Overall, the Wald test is much more powerful than the Hommel method. Two empirical examples, in the EG and NEAT designs, are used, showing that the Wald test can provide different conclusions to other methods in practice.

4. Conclusions

In the past ten years, kernel equating has been developed on many levels. To mention only a few developments, new kernels have been included (Lee and von Davier, 2011), the utility of the standard error of equating difference has been investigated (Moses and Zhang, 2011), new ways to assess the statistical significance of differences between equating methods have been proposed (Rijmen et al., 2011) and local equating methods have been introduced (Wiberg et al., 2014).

This thesis contributes to the kernel method of equating in several important ways. Kernel equating now incorporates all common equating designs for tests consisting of dichotomous items and an easily accessible software implementation for all designs has been created. With these additions, the kernel method of test equating is perhaps the most comprehensive and easily used observed-score equating method for practitioners. The bandwidth selection in kernel equating has been improved by providing an alternative data-driven way to select the bandwidth parameters when the data are not smooth and by accounting for the bandwidth selection when estimating the standard errors. In addition to IRT observed-score equating being included in the kernel equating framework, the method has also been generalized to the multivariate case. This generalization allows for hypothesis testing of equating differences across more score points than previously possible. Hypothesis testing of equating differences has been further investigated and shown to offer a powerful method to detect differences between NEAT CE and NEAT PSE equating functions. There now exist more methods to choose from when conducting an equating and the new hypothesis testing methods will help in applied work when determining which equating function should be used for a given test administration.

There are many possible future research topics in the area of kernel equating. When tests consist of items scored in multiple categories, so called polytomous items, the asymptotic covariance matrix of the resulting IRT observed-score equating function can be derived. Additional equating coefficient estimators for the IRT NEAT PSE equating method can be integrated in the kernel equating framework and the results compared to current methods. Another useful addition to the kernel equating framework would be to derive the asymptotic results when taking the bandwidth selection with penalty functions into account.

Acknowledgements

First, I want to thank my advisor Fan Yang-Wallentin for believing and placing trust in me. The resilience you have shown during the past year is incredible. I also want to thank my assistant advisors Marie Wiberg and Alina A. von Davier. You have introduced me to new research areas and have given me opportunities and experiences I never would have imagined I could get.

I wish to thank Ingeborg Waernbaum for introducing me to the topic of graphical models and for recommending me to pursue a PhD.

The Faculty of Social Sciences at Uppsala University has been supportive of me as a PhD student for which I am very thankful.

I want to thank all my co-workers at the department. A special thanks to Bo Wallentin for his recommendation to start working on research early in the PhD program.

Shelby Haberman, Hongwen Guo, Tim Moses, Rolf Larsson, Anton Béguin, Inger Persson, Rauf Ahmad, Johan Lyhagen and many anonymous referees are thanked for their numerous comments on different parts of the thesis.

I would also like to extend a thank you to all the PhD students at the department. In particular, I want to thank my office mate 金少博 for answering my elementary statistical queries and for being a good friend. The knowledge I gained from the courses in the PhD program was enhanced by discussions with David Kreiberg. Also, Xingwu Zhou is thanked for the many amusing moments he provided.

I want to thank my friends, notably Susanne Asp and Linda Korol. You have kept my spirits up throughout the PhD program. 罗昊 is thanked for lighting up my days the past year.

Lastly I want to thank my family, especially my mother who has always encouraged me to stand up for what I believe in.

References

Angoff, W. H. (1984). *Scales, Norms and Equivalent Scores*. Princeton, NJ: Educational Testing Service. (Reprinted from Thorndike, R. L. (Ed.) *Educational Measurement*, 1971).

Bränberg, K. and Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, 41:419–440.

Brennan, R., Wang, T., Kim, S., and Seol, J. (2009). Equating recipes [computer program]. *Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa*.

Chen, H., Yan, D., Hemat, L., Han, N., and von Davier, A. A. (2011). *LOGLIN/KE User Guide*. Princeton, NJ: Educational Testing Service. Version 3.1.

Dorans, N. J. and Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37:281–306.

Ferguson, T. (1996). *A Course in Large Sample Theory*. Chapman & Hall texts in statistical science series. London: Chapman & Hall.

Hambleton, R. K. and Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Holland, P., King, B. F., and Thayer, D. T. (1989). The standard error of equating for the kernel method of equating score distributions. Technical Report 89-83, Princeton, NJ: Educational Testing Service.

Holland, P. W. and Thayer, D. T. (1989). The kernel method of equating score distributions. Technical Report 89-84, Princeton, NJ: Educational Testing Service.

Holland, P. W. and Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, 25(2):133–183.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383–386.

Kolen, M. J. and Brennan, R. J. (2014). *Test Equating: Methods and Practices (3rd ed.)*. New York, NY: Springer-Verlag.

Lee, Y.-H. and von Davier, A. A. (2011). Equating through alternative kernels. In von Davier, A. A., editor, *Statistical Models for Test Equating, Scaling, and Linking*. New York, NY: Springer-Verlag.

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

Lord, F. M. and Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, 8:452–461.

Moses, T. and Zhang, W. (2011). Standard errors of equating differences: Prior developments, extensions, and simulations. *Journal of Educational and Behavioral Statistics*, 36:779–803.

Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika*, 68:193–211.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rijmen, F., Qu, Y., and von Davier, A. A. (2011). Hypothesis testing of equating differences in the kernel equating framework. In von Davier, A. A., editor, *Statistical Models for Test Equating, Scaling, and Linking*, Statistics for Social and Behavioral Sciences, pages 317–326. New York, NY: Springer.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, 17(5):1–25.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. New York, NY: Chapman and Hall/CRC.

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, 78:605–623.

von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). *The Kernel Method of Test Equating*. New York, NY: Springer-Verlag.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482.

Wiberg, M., van der Linden, W. J., and von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement*, 51:57–74.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 106

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-234618

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2014