Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 29 Number 5, March 2024 ISSN 1531-7714

Impacts of Differences in Group Abilities and Anchor Test Features on Three Non-IRT Test Equating Methods[1]

Inga Laukaityte, *Umeå University*, Marie Wiberg, *Umeå University*

The overall aim was to examine effects of differences in group ability and features of the anchor test form on equating bias and the standard error of equating (SEE) using both real and simulated data. Chained kernel equating, Postratification kernel equating, and Circle-arc equating were studied. A college admissions test with four different anchor test forms administered at three test administrations was used. The simulation study examined the differences in ability of the test groups, and differences in the anchor test form with respect to item difficulty and discrimination. In the empirical study, the equated values from the three methods only slightly differed. The simulation study indicated that an easier anchor test form and/or an easier regular test form, and anchor items with a wider spread in difficulty, negatively affected the SEE and bias. The ability level of groups was also important. Equating with only less or more capable groups resulted in high SEEs at higher and lower test scores, respectively. The discussion includes practical recommendations to whom an anchor test should be given if there is a choice and how to select an anchor test form which have equating as primary purpose.

Keywords: NEAT, chained kernel equating, Postratification kernel equating, Circle-arc equating, admission test, high stakes assessment

Introduction

When equating test scores, we use statistical models and methods to map test scores from one scale to their equivalents on another scale, so that the test scores can be used interchangeably (Gonzaléz & Wiberg, 2017). In test score equating it is very useful to use a nonequivalent group with an anchor test (NEAT) design, where there are two samples from two nonequivalent populations, who take different test forms and a common anchor test form. In a NEAT design, the ability of the populations might differ substantially as the test takers may take the test forms at different time points. This is problematic as the quality of the equating is affected by differences in the groups' ability (Cook & Petersen, 1987; Kolen, 1990; Liu, Sinharay, Holland, Curley, & Feigenbaum, 2011; Sinharay & Holland, 2007).

To examine the ability of different groups, an anchor test form can be used. Early research on the construction of an anchor test form suggested that a minitest should be used, i.e., a miniature version of the total test with respect to both content and statistical characteristics (Angoff, 1968; Klein & Jarjoura, 1985; Kolen & Brennan, 2014; Petersen, Marco & Stewart, 1982; Petersen, Kolen & Hoover, 1989, p. 246; von Davier, Holland & Thayer, 2004, p. 33). This essentially means matching content and the item difficulty distribution in terms of mean and variance and the average item discrimination (Dorans, Kubiak & Melican, 1998). Subsequently, Sinharay and Holland

[1] Funding information: The research was partially funded by the Swedish Wallenberg MMW 2019.0129 grant.

(2006a; 2007) suggested that anchor test forms could be used with smaller variances of item difficulties, so the item difficulty variance of the total test and anchor test form do not need to be matched. Their suggestion to use a miditest, i.e., an anchor test with items of medium difficulty, led to better equating compared to the minitest in some practical situations. Their suggestion was confirmed with SAT data when comparing a miditest and a minitest (Liu et al, 2011; Liu, Sinharay, Holland, Feigenbaum, & Curley, 2011). In these studies, item response theory (IRT) true-score equating was used together with chained kernel equating and poststratification equating. These methods were chosen as they are commonly used when equating with a NEAT design and in operational settings of the SAT. Sinharay, Haberman, Holland and Lewis (2012) also proposed that a miditest was most suitable for an anchor test. Trierweiler, Lewis, and Smith (2016) found that making an anchor test a miditest does not generally maximize the anchor test to total test correlation. However, Sinharay (2018) showed that both a minitest and a miditest are potentially useful as some of the situations described by Trierwailer et al. (2016) are not realistic. In addition, the shorter the length of the anchor test form, the larger the bias, especially for linear kernel and chained kernel equating compared with frequency estimation and kernel poststratification equating (Ricker & von Davier, 2007). Sinharay and Holland (2006b) also pointed out that large population differences appear to have the largest impact on equating in terms of root mean squared error. Puhan (2010) examined different linear equating methods under different test conditions with respect to root mean squared error and bias. A conclusion was that either chained linear or Levine equating should be used when the samples of test takers who took compared test forms differed in ability. Powers and Kolen (2014) examined group differences when using frequency estimation, chained equating, and IRT observed-score and IRT true-score equating. Their study was based on real test data, and they concluded that IRT equating and chained equating was less sensitive to group differences compared with frequency estimation. Despite the previous studies summarized here, it is still not clear how different anchor test forms in terms of discrimination and difficulty affect the equating transformations when there are group ability differences. Liu et al (2011) suggested, from studying empirical data, that there is an interaction between the type of anchor test (minitest or miditest) and group ability difference. The present study differs from previous studies as it both includes real and simulated data, where most previous studies only include real data. Our study also focuses on Circle-arc and kernel equating methods, which have not been previously examined in this context. A reason for these choices is that primary classical test theory (CTT) sum scores, and not IRT, are applied in operational settings when reporting and equating test scores of the test we studied. The test we used is the Swedish Scholastic Aptitude Test (SweSAT), a college admissions test administered at several test centers around Sweden. The abilities and score distributions of the test takers connected to these centers are known to differ as some test centers are in university cities while others are in labor cities. As anchor test forms are only given to a relatively small number of test takers, a choice must be made concerning who should receive the anchor test form. In other words, to which test taker group in terms of ability is it better to give an anchor test form to minimize standard errors and bias. To the best of our knowledge, there have been no previous analyses, or we have not seen any study on who should be given an anchor test form when one knows that the test taker groups ability differ. Especially when the test is based on CTT sum scores.

The overall aim of our study was to examine how differences in groups' abilities and various features of the anchor test form affect the equating transformation and standard error of equating (SEE) using both real data and by varying several conditions in a simulation study. Chained kernel equating, Postratification kernel equating, and Circle-arc equating were used as these methods are applied when equating the examined college admissions test forms and have yielded stable equating results in such settings in the past. The examined conditions included differences in the anchor test forms with respect to item difficulty and item discrimination, and differences in abilities of the groups who received the anchor test form. Note, our focus is on the statistical properties of the anchor test form and not on its content, because in the empirical study we used anchor test forms with similar content to the regular test forms. Also, as noted above, our study is set in a CTT context as SweSAT scores and equating are theoretically rooted in CTT (Wedman & Lyrén, 2015; Lyrén, 2009).

The rest of this paper is structured as follows. The next section describes how to equate test scores using

a NEAT design with the three methods of interest. This is followed by a section presenting an empirical study of a real college admission test and its results, then a corresponding simulation study. The paper ends with a discussion, including practical recommendations and suggestions for further research.

Equating test scores using the NEAT design

Assume that we have a test form X with test scores X that form a random variable from population **P** and another test form Y with test scores Y which is a random variable from population **Q**. In the NEAT design we assume that we have access to an anchor test form, comprised of several common items. The anchor test form can be used to estimate the difficulty level of the test forms and test takers' ability. To find a score y on test form Y equivalent to a score x on test form X we assume that X and Y are continuous, and their cumulative density functions (CDFs) are denoted $F_X(x)$ and $G_Y(y)$, respectively. In general, the equipercentile equating transformation can then be defined as:

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

As stated in the introduction, we use CTT as the underlying measurement model here. There are numerous available equating methods including IRT equating methods that can be used with a NEAT design. However, we did not use IRT based methods as CTT methods have typically been used for equating test forms of the admissions test used in the empirical study. Hence, we used Circle-arc equating, chained kernel equating and poststratification kernel equating in both the empirical and simulation studies. As already mentioned, these three equating methods are used operationally when equating the college admissions test and have yielded stable equating results over administrations. Note, one of the equating requirements is that it should be population invariant, as discussed for example by Kolen and Brennan (2014), so we can use a subsample for the equating. In this study, we will examine population invariance by assessing the impact of various group differences on equating outcomes using the three chosen methods.

Kernel equating (von Davier, Holland & Thayer, 2004) involves five steps. Step 1 is *presmoothing* the observed test score distributions, typically using loglinear modeling, and step 2 *estimation of score probabilities* from the models selected in step 1. Step 3 is *continuization* of the discrete test score distributions obtained from step 2, which includes use of a continuous random variable that characterizes the selected kernel (for example, uniform, logistic or Gaussian). We used the Gaussian kernel in this study as it is applied in operational settings of the SweSAT sources of data used in our empirical study. The obtained continuized CDFs of test scores X and Y are respectively defined as $F_{h_X}(x)$ and $G_{h_Y}(y)$, where H_x and h_Y are the bandwidths that control the degree of smoothness in the continuization. Note, the bandwidths can be selected in several ways (Häggström & Wiberg, 2014), but the methods tend to give similar equating results (Wallin, Häggström & Wiberg, 2021). The penalty function described extensively by von Davier et al. (2004) was used both in the empirical and in the simulation study as it is a common choice. Briefly, using a penalty function we want to find the bandwidth h_X, by minimizing $\text{PEN}(h_X) = \sum_j (\hat{r}_j - \hat{f}_{h_X}(x_j))^2 + \kappa \cdot \sum_j A_j$ where $\hat{r}_j$ is the estimated score probabilities, $\hat{f}_{h_X}(x_j)$ is the estimated density function of the continuous transformation of X. κ is a constant set to 1 if the second penalty term is used, and otherwise 0. $A_j = 1$ if the derivatives are smooth, and otherwise 0. Step 4 is the actual *equating*. In a NEAT design, one option is to use Poststratification equating (PSE), in which we first condition X and Y on A then reweight their distributions to estimate the CDFs. PSE assumes that the conditional distribution of X given A and the conditional distribution of Y given A, are the same in populations P and Q. The equating is done on the synthetic population, which is a combination of populations P and Q. The kernel poststratification equating (KPSE) is defined as

$$\varphi_Y(x) = G_{h_Y}^{-1}(F_{h_X}(x)) \quad (2)$$

The other alternative is to use chained equating (CE), in which test form X is linked to the anchor test form A, which is then linked to test form Y in a chain. We used kernel chained equating (KCE), defined as:

$$\varphi_Y(x) = G_{h_Y}^{-1}(H_{h_Y}(H_{h_X}^{-1})(F_{h_X})(x))) \quad (3)$$

where H_{h_Y} and H_{h_X} are the continuized CDFs for the anchor test scores from the anchor test forms given to the groups that received test forms X and Y, respectively. Previous studies (e.g., Kolen & Brennan,

2014; Sinharay & Holland, 2007; Liu, Sinharay, Holland, Curley, & Feigenbaum, 2011; Livingston et al., 1990) have concluded that traditional frequency estimation (i.e., PSE but not within the kernel equating framework) is more sensitive than traditional CE when groups are known to differ. This is because the population invariance assumption for PSE methods is not likely to hold when groups differ substantially (Holland & Dorans, 2006; von Davier et al., 2004). Livingston, Dorans, and Wright (1990) proposed that chained equipercentile equating should be considered when there is substantial difference between the groups. Further, von Davier et al. (2004b) concluded that chained equating and frequency estimation should give the same results if the populations are equivalent or if the anchor and total test scores are perfectly correlated. Their results also imply that these methods are likely to yield different results in practice when there are large between-group differences. Several authors (e.g., Wang, Lee, Brennan, & Kolen, 2008; Sinharay & Holland 2010a, b; Sinharay et al. 2011) have concluded that if there are large differences between the groups, traditional frequency estimation tends to have larger bias but smaller SEE than traditional chained equating. It should be noted that we have not found any corresponding analyses of kernel equating but we expect kernel chained equating and kernel poststratification equating to behave similar to chained equating and frequency estimation. In the fifth step, *evaluating the equating transformation*, different accuracy measures are calculated, including, for example, bias and SEE values (Wiberg & González, 2016).

Circle-arc equating is a classical, nonlinear equating transformation based on a method presented by Divgi (1987), in which an equating curve is constrained to pass through three points: two prespecified endpoints and a middle point. The lower endpoint (x_1, y_1) is determined by the lowest meaningful score which, in our (multiple-choice test) case, is the chance score (guessing) on each test form. The upper endpoint (x_3, y_3) is determined by the highest meaningful score on each test form. The middle point (x_2, y_2) is determined empirically, for example, from the mean scores for each test form. The three points are connected by drawing an arc.

In the simplified approach to the Circle-arc method – Method 2 described by Livingston and Kim (2009) – the equating function combines the linear component defined by the low and high points,

$$L(x) = y_1 + \frac{y_3 - y_1}{x_3 - x_1}(x - x_1),$$

and a curvilinear component defined by the center coordinates (x_c, y_c) and radius *r* of the circle:

$$y^* = y_c \pm \sqrt{r^2 - (x - x_c)^2}.$$

The Circle-arc equating function is the sum of the linear and curvilinear components. We are aware that Circle-arc equating is typically used for small samples but over the years it has been used successfully as one of the equating methods in SweSAT. It should be noted that Circle-arc equating does not directly address differences in groups' abilities, and we have found no research studies where this has been examined.

To evaluate the equating transformations, we compared the equated values to the original scores then examined the SEE, and (in the simulation study) the bias. The reason for using SEE is that it has worked well in previous equating studies, e.g., Wallin, Häggström and Wiberg (2021), and is used in operational settings in the college admissions test of the empirical study. von Davier et al. (2004) provide a detailed description of how to obtain SEE in kernel equating, but briefly it measures the uncertainty in an equating transformation, and for equating *X* to *Y* is generally defined as:

$$\text{SEE}_Y = \sqrt{\text{Var}(\hat{\varphi}_Y(x))}$$

In the empirical study, SEE values for Circle-arc method were obtained by calculating the bootstrap standard errors.

Let $\hat{\varphi}_Y(x_i)$ denote the equating estimator evaluated at a particular score x_i on a test form X using sample data, and $\varphi_Y(x_i)$ denote the population equating transformation, then bias (a measure of systematic equating error) at score x_i for a given equating can be defined as:

$$\text{Bias}[\hat{\varphi}_Y(x_i)] = \hat{\varphi}_Y(x_i) - \varphi_Y(x_i).$$

In the simulation study with R replications the bias for test score x_i was defined as:

$$\text{Bias}[\hat{\varphi}_Y(x_i)] = \frac{1}{R}\sum_{g=1}^{R}[\hat{\varphi}_Y^{(g)}(x_i) - \frac{1}{R}\sum_{g=1}^{R}\varphi_Y^{(g)}(x_i)]$$

Test forms X and Y were always simulated to be equivalent, thus the linear equating function was

considered as a valid choice for the criterion function $\varphi_Y(x_i)$. Linear equating function has also been used as a criterion function for example in Lieu et al. (2011). Note, in all simulated conditions, linear equating function appeared to be the same as the identity function.

We also used the weighted absolute bias (WAB), as described for example by Lieu et al. (2011), which is a summation of the differences at each score point and calculated as:

$$\text{WAB}[\hat{\varphi}_Y(x)] = \frac{1}{N}\sum f_{x_i} |\text{Bias}[\hat{\varphi}_Y(x_i)]|,$$

where N is the number of test takers who received the new test form and were part of the equating sample. In the empirical study N varied between the samples while in the simulation study we used $N = 2{,}000$. f_{x_i} is the frequency at a particular score x_i in the new test form group.

Note, the choice of an equating method should align with the equating process's goals. If minimizing bias is the primary concern, equating methods that yield results with the lowest bias should be prioritized. On the other hand, if reducing SEE or improving the precision of equating is the primary goal, then methods that minimize the SEE and maximize precision should be chosen. In practice, there is often a trade-off between bias and SEE. Some equating methods excel at reducing bias but yield higher SEE, while others may minimize SEE but introduce some bias. The specific context, the consequences of bias or imprecision in the equating process, and the stakeholders' priorities should all be considered.

In both the empirical and simulation studies the R package *kequate* (Andersson, Bränberg & Wiberg, 2013) was used for all the kernel equating and the R package *equate* (Albano, 2016) for the Circle-arc equating. All codes used can be obtained from the corresponding author upon request.

Empirical Study

For our empirical comparisons we used scores obtained in the SweSAT college admissions test, which is typically administered twice a year, once in spring (labeled A) and once in fall (labeled B). The SweSAT form includes 160 dichotomously scored multiple choice items divided into a verbal section and a quantitative section, each including 80 items, and are separately equated. Each section is given to the test takers as a test part comprising 40 items. Thus, each test taker is given two verbal test parts and two quantitative test parts, designed (content-wise) as minitests. The total SweSAT score is the sum of the scores obtained on the verbal and quantitative sections, and ranges from 0 to 160. The test takers also receive an extra test part of 40 items, which can be either verbal or quantitative, and contains either try-out items or an external anchor test form. The test takers do not know which test part is a regular test part, external anchor test form or try-out part. Thus, the test takers are administrated a total of 200 items distributed equally in five parts.

We had access to scores from three administrations of the SweSAT (2016B, 2017A and 2018A) and two 40-item verbal anchor test forms (labeled V1 and V2) and two quantitative (referred to as *quant*) anchor test forms (labeled Q1 and Q2). Note that each of the verbal and quant anchor test forms were given to different groups of test takers so they received either a verbal or a quantitative anchor test form, or a verbal or quantitative try-out item part. The anchor test forms we used in this study have been used for several years for equating purposes. The form used in each SweSAT administration (which we subsequently refer to as a regular test form) was received by between 40,000 and 75,000 test takers, but only around 2,000 test takers received an anchor test form to maintain test security. Therefore, equating was limited to partial datasets from one or two test centers that provided both anchor test and the regular test scores for test takers. (These partial data sets of regular test forms are labeled Reg Q1, Reg Q2, Reg V1 and Reg V2, depending on which anchor test form (Q1, Q2, V1 and V2) the test taker also answered). This means that one can select one or two test centers from the 21 available test centers across Sweden for administering the anchor test form. The average abilities of test-takers from these centers are known from previous years and show significant variations. Thus, it is possible that abilities of the groups that received the anchor test form may have influenced the equating transformation as CTT rather than IRT is used as the measurement model (because CTT underlies SweSAT, and sum scores are used to facilitate the public's understanding of SweSAT scores).

The anchor test forms were examined using descriptive statistics including correlations, difficulty, and the shape of the test score distributions. The difficulty of the anchor test forms was measured through the overall mean item difficulties, defined as the proportions of test takers who correctly answered the items. Thus, a lower value on the mean anchor test form difficulty indicates a more difficult test form. We examined means and standard deviations of item difficulties and item discrimination, and the item parameters for evidence of parameter drift using the Mantel-Haenszel statistic. We also assessed the items and test forms with IRT using the R package *mirt* (Chalmers, 2012) to get realistic estimates of the item parameters that we could use in the later simulation study. Note, IRT is only used as an evaluation tool of the SweSAT items, but the test forms are neither scored nor equated with IRT.

We assessed effects of differences in anchor test forms (Q1 versus Q2 and V1 versus V2) on the equating transformation in two scenarios. In Scenario 1 we equated test form 2018A to test form 2016B (to equate the form used in a recent administration to one used in the oldest administration). In Scenario 2 we equated test form 2018A to test form 2017A (to equate a form used in the same recent administration to the one used in the closest administration at the same time of the year). As most students who plan to study at university in the fall take the spring administration A, we chose that administration as the base test form which we equated from. We examined the two scenarios with KCE, KPSE and Circle-arc. For the two kernel equating methods (KCE and KPSE), we used a quadratic (second order) polynomial model with one interaction term as a presmoothing model for the NEAT design as we followed the parsimony principle, and these models showed a good fit. The weight in KPSE was set to 0.50 as the group sizes were similar in all used samples. The equating transformations were evaluated with the SEE.

Results of the Empirical Study

Descriptive statistics

Descriptive statistics of the item difficulties of the regular test forms and the anchor test forms are presented in Table 1. The mean difficulties of anchor test forms Q2 and V2 were comparable to those of the regular test forms, but the anchor test forms Q1 and V1 were more difficult, on average, than the regular test forms. The standard deviations of the item difficulty of the anchor forms were similar to those of the regular test forms. The results indicate that anchor test form V1 had the widest range of item difficulties. Note, there was no sign of item parameter drift, so results of its assessment are not included, but can be provided upon request.

Raw score descriptive statistics for the anchor and regular test forms are shown in Table 2. The correlation between the regular test forms (2016B, 2017A, and 2018A) and anchor test forms varied from 0.83 (for Q1 in 2016B) to 0.88 (for Q2 in 2017A and 2018A). The mean raw scores varied from 17.51

Table 1. Descriptive statistics of the regular test and anchor test forms' item difficulties (p-values)

	Items	Mean difficulty	SD of difficulty	Min	Max		Items	Mean difficulty	SD of difficulty	Min	Max
2016B Quant						**2016B Verbal**					
Reg Q1	80	0.56	0.15	0.25	0.85	Reg V1	80	0.52	0.13	0.27	0.79
Q1	40	0.44	0.12	0.22	0.76	V1	40	0.44	0.17	0.10	0.82
Reg Q2	80	0.53	0.15	0.22	0.83	Reg V2	80	0.57	0.14	0.33	0.83
Q2	40	0.49	0.14	0.26	0.74	V2	40	0.56	0.15	0.25	0.83
2017A Quant						**2017A Verbal**					
Reg Q1	80	0.54	0.12	0.28	0.76	Reg V1	80	0.56	0.15	0.23	0.82
Q1	40	0.45	0.13	0.20	0.75	V1	40	0.48	0.17	0.11	0.84
Reg Q2	80	0.54	0.12	0.28	0.77	Reg V2	80	0.55	0.15	0.23	0.82
Q2	40	0.53	0.15	0.26	0.82	V2	40	0.54	0.15	0.23	0.84
2018A Quant						**2018A Verbal**					
Reg Q1	80	0.51	0.12	0.25	0.77	Reg V1	80	0.53	0.14	0.23	0.83
Q1	40	0.44	0.13	0.21	0.75	V1	40	0.45	0.18	0.09	0.81
Reg Q2	80	0.53	0.13	0.23	0.80	Reg V2	80	0.55	0.15	0.27	0.86
Q2	40	0.54	0.15	0.26	0.81	V2	40	0.58	0.15	0.25	0.88

(2018A) to 17.89 (2017A) for the groups that received the quantitative anchor forms Q1, and from 19.55 (2016B) to 21.41 (2018A) for Q2. For the groups that received the verbal anchor test forms, the mean raw score varied from 17.51 (2016B) to 19.01 (2017A) for the group that received V1, and from 21.69 (2017A) to 23.06 for the group that received V2 (2018A).

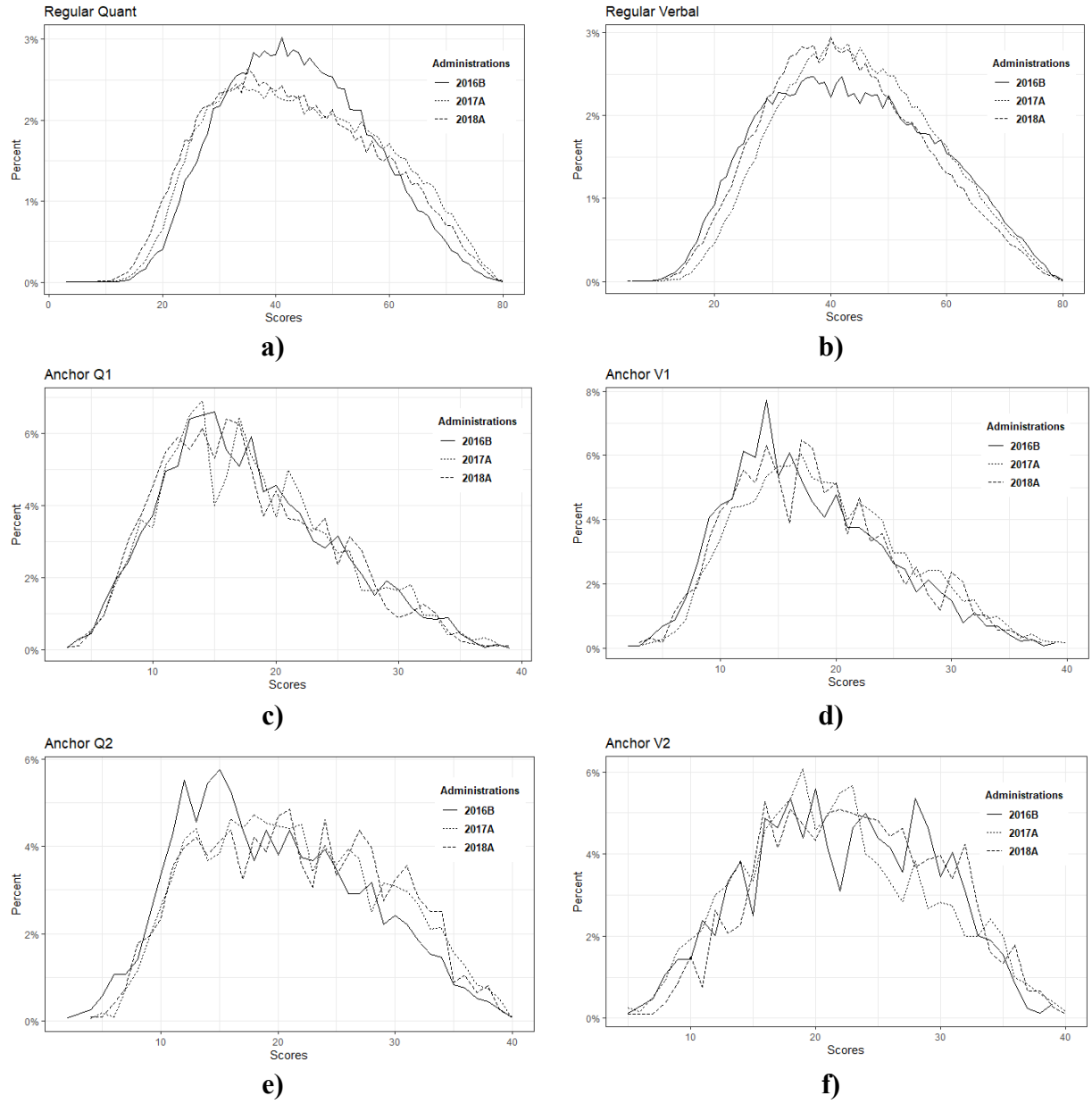
We let P represent those who took the old test forms (2016B or 2017A, depending on the comparison) and Q represent the group who took the new form (2018A). To test the groups' equivalence, we calculated the effect size of the mean in terms of the standardized mean difference (SMD): $\frac{\bar{A}_P - \bar{A}_Q}{SD_{A(P+Q)}}$, where $\bar{A}_P$ and $\bar{A}_Q$ are mean scores of populations P and Q for the anchor form, and $SD_{A(P+Q)}$ is the standard deviation of the anchor score for the combined population $P + Q$. More details can be obtained from works by authors including Liu, Sinharay, Holland, Curley, & Feigenbaum, 2011. As can be seen from Table 2, all absolute SMD values were within the interval 0.2-0.5. SMD was -0.03 (old regular test form 2016B) and -0.06 (old regular test form 2017A) on the anchor test form Q1, indicating that the groups who took the old and new test forms were similar or close to identical at their ability levels. Similarly, SMD was 0.10 (old test form 2016B) and -0.12 (old test form 2017A) on the anchor test form V1, indicating that there were small differences in ability between the groups who took these tests. However, there was a substantial difference in the SMDs obtained in the two equating scenarios designed to assess effects of differences in anchor test forms. For anchor test form Q2 the SMD was 0.24 in Scenario 1 and close to zero (0.01) in Scenario 2. Differences in SMDs between the two scenarios were also obtained for the verbal anchor test form V2, with SMD values of 0.12 and 0.19 in Scenarios 1 and 2, respectively.

The test score distributions for each of the administrations are presented in Figure 1. The distributions of scores for the anchor test forms

Table 2. Descriptive statistics of raw scores for the anchor and regular test forms of the three included administrations (Adm). SMD is the Standardized Mean Difference.

Adm	Test form	Items	Sample size	Mean	SD	Skewness	Kurtosis	Correlation	SMD: new-old (anchor) forms
2016B	Reg Q1	80	2439	44.63	12.27	0.18	-0.59	0.83	
2016B	Q1	40	2439	17.72	6.81	0.50	-0.28		-0.03
2016B	Reg Q2	80	1578	42.33	12.23	0.35	-0.45	0.84	
2016B	Q2	40	1578	19.55	7.59	0.34	-0.66		0.24
2017A	Reg Q1	80	1445	43.44	13.98	0.32	-0.80	0.85	
2017A	Q1	40	1445	17.89	6.86	0.49	-0.32		-0.06
2017A	Reg Q2	80	2288	43.48	13.98	0.26	-0.80	0.88	
2017A	Q2	40	2288	21.34	7.69	0.22	-0.83		0.01
2018A	Reg Q1	80	1564	41.18	14.17	0.40	-0.65	0.86	
2018A	Q1	40	1564	17.51	6.76	0.52	-0.30		-
2018A	Reg Q2	80	1237	42.03	13.82	0.24	-0.77	0.88	
2018A	Q2	40	1237	21.41	7.76	0.11	-0.92		-
2016B	Reg V1	80	1889	41.88	13.82	0.28	-0.76	0.85	
2016B	V1	40	1889	17.51	6.80	0.54	-0.22		0.10
2016B	Reg V2	80	841	45.68	13.35	-0.02	-0.80	0.87	
2016B	V2	40	841	22.24	7.07	-0.20	-0.82		0.12
2017A	Reg V1	80	1855	44.95	12.29	0.25	-0.49	0.85	
2017A	V1	40	1855	19.01	6.91	0.41	-0.38		-0.12
2017A	Reg V2	80	1201	44.23	12.49	0.28	-0.56	0.85	
2017A	V2	40	1201	21.69	7.30	0.22	-0.62		0.19
2018A	Reg V1	80	1266	42.22	13.01	0.30	-0.47	0.86	
2018A	V1	40	1266	18.16	6.90	0.46	0.07		-
2018A	Reg V2	80	1061	43.96	12.11	0.16	-0.57	0.86	
2018A	V2	40	1061	23.06	6.91	0.04	-0.71		-

Figure 1. Test score distributions for the following forms: a) regular Quant, b) regular Verbal, c) anchor Q1, d) anchor V1, e) anchor Q2, and f) anchor V2.



seem to be either more positively skewed (Q1, V1) or flatter (Q2, V2) than those of the regular test forms.

To explore the different anchor test forms more thoroughly and enable subsequent set-up of a realistic simulation study, we examined them using the three-parameter logistic IRT model with concurrent calibration, so all parameters were on the same scale. The overall results indicated that abilities of the groups that received these four anchor test forms were quite similar. The abilities of the test takers varied from -3.04 to 3.34. The overall mean ability, calculated across the three administrations that were analyzed, was 0.03 for the groups that received either the quant anchor test form Q1 or Q2. When considering each administration individually, the mean ability ranged from -0.12 to 0.10 for the quant anchor test forms. Similarly, the overall mean ability was 0.02 for the groups that took the verbal anchor test forms V1 or V2. When considering each administration individually, the mean ability ranged from -0.07 to 0.13 for the verbal anchor test forms. Note, although the mean abilities were quite similar in these samples, the mean ability on different test centers, during the period the test results of test

takers were valid, varied between -0.11 and 0.29 in other administrations. The item discrimination parameters varied most for anchor test form V1, from 0.31 to 2.84 (with mean 1.43), while they varied from 0.71 to 2.09 (mean 1.26) for V2. For the Quant part, item discrimination parameters varied from 0.45 to 2.77, with means of 1.34 for Q1 and 1.63 for Q2. Mean item difficulties were highest for the anchor test forms Q1, 0.83 ([-1.50;2.68]) and V1, 0.64 ([-1.46;2.71]). Mean item difficulties for anchor test forms Q2 (0.46 [-1.91;2.48]) and V2 (0.27 [-1.61;2.21]) were half those of anchor test forms Q1 and V1, respectively. The mean lower asymptote parameters for anchor test forms Q1 (0.17 [0.01;0.42]) and V1 (0.15 [0.01;0.38]) were lower and less spread than those of anchor test forms Q2 (0.21 [0.02;0.63]) and V2 (0.21 [0.004;0.44]). Results of the item parameter distributions were used in the simulation study, and complete results of the IRT analyses can be obtained from the corresponding author upon request. Figure 2 shows the test information functions for each test form.

Equating results

Figure 3 shows results of the equating transformations for the quant anchor test forms (Q1 and Q2, left column) and verbal anchor test forms (V1 and V2, right column) obtained with the three equating methods in Scenarios 1 (upper row) and 2 (lower row). In scenario 1, there were small differences in the equated values both between the equating methods and between anchor test forms. For the quant test forms these small differences were observed in both the lower and upper score ranges, while for the verbal test forms small differences were only observed in the lower score range. Note that KPSE and KCE produced similar equating results for the verbal anchor test forms (Fig. b). In scenario 2 there were only small differences between the equating methods, and differences were largest in the lower score range.

Figure 4 shows SEE values obtained for the four anchor test forms with the three equating methods in Scenarios 1 (upper row) and 2 (lower row). The highest SEEs using the quant test forms were obtained with the KPSE and KCE equating methods when using anchor test form Q1 in Scenario 1, while for the verbal anchor test forms the highest SEEs were obtained with anchor test form V2 in Scenario 2. The lowest SEEs were obtained with Circle-arc equating in both scenarios and for all anchor test forms.

In summary, all considered groups of test takers had similar abilities and the equating transformations only differed slightly, especially in the lower score range (Figure 3) and more in Scenario 1 than in Scenario 2. The lowest SEE values were obtained with Circle-arc equating in both scenarios. They were also lower in the mid-score range than in the lower and upper score ranges when using KPSE and KCE (but even in this range Circle-arc equating yielded lower values) and KPSE yielded smaller values than KCE (Figure 4).

The differences between results for the two scenarios were generally larger than those obtained using different anchor test forms. As we cannot generalize the findings from an empirical example, we also conducted a simulation study.

Statistical Analyses

We conducted a simulation study to examine impacts of the anchor test features and differences in ability between the anchor test groups on the equating transformation. To connect this study with the SweSAT data we decided to mirror the set up by using regular test forms with 80 items and anchor test forms with 40 items. We used the estimated IRT parameters of the regular and anchor test forms from the empirical study to choose suitable distributions for the corresponding forms in the simulation study. As SweSAT only includes multiple choice items, we used the three-parameter logistic IRT model to generate anchor test forms and regular test forms X and Y.

Simulated data and examined conditions

We constructed 23 scenarios, summarized in Table 3. Those with similar conditions have been designated with the same number but different letters. We generated 2,000 test takers who took each test form, as in the real test setting. Group *P* took test form X and group *Q* took test form Y. The first scenario (S1) is the baseline case, assuming that the groups who took the two test forms had similar ability level (as seen in the empirical study), set to $\theta \sim N(0,1)$ for both groups from populations P and Q. However, as the mean ability of the test takers at different test centers are known to differ, we examined effects of one group being More (+) or less (-) capable than the other by adding or subtracting 0.5, respectively, to the mean value of the

Figure 2. Test information plots for the following forms: a) regular Quant, b) regular Verbal, c) anchor Q1, d) anchor V1, e) anchor Q2, and f) anchor V2.

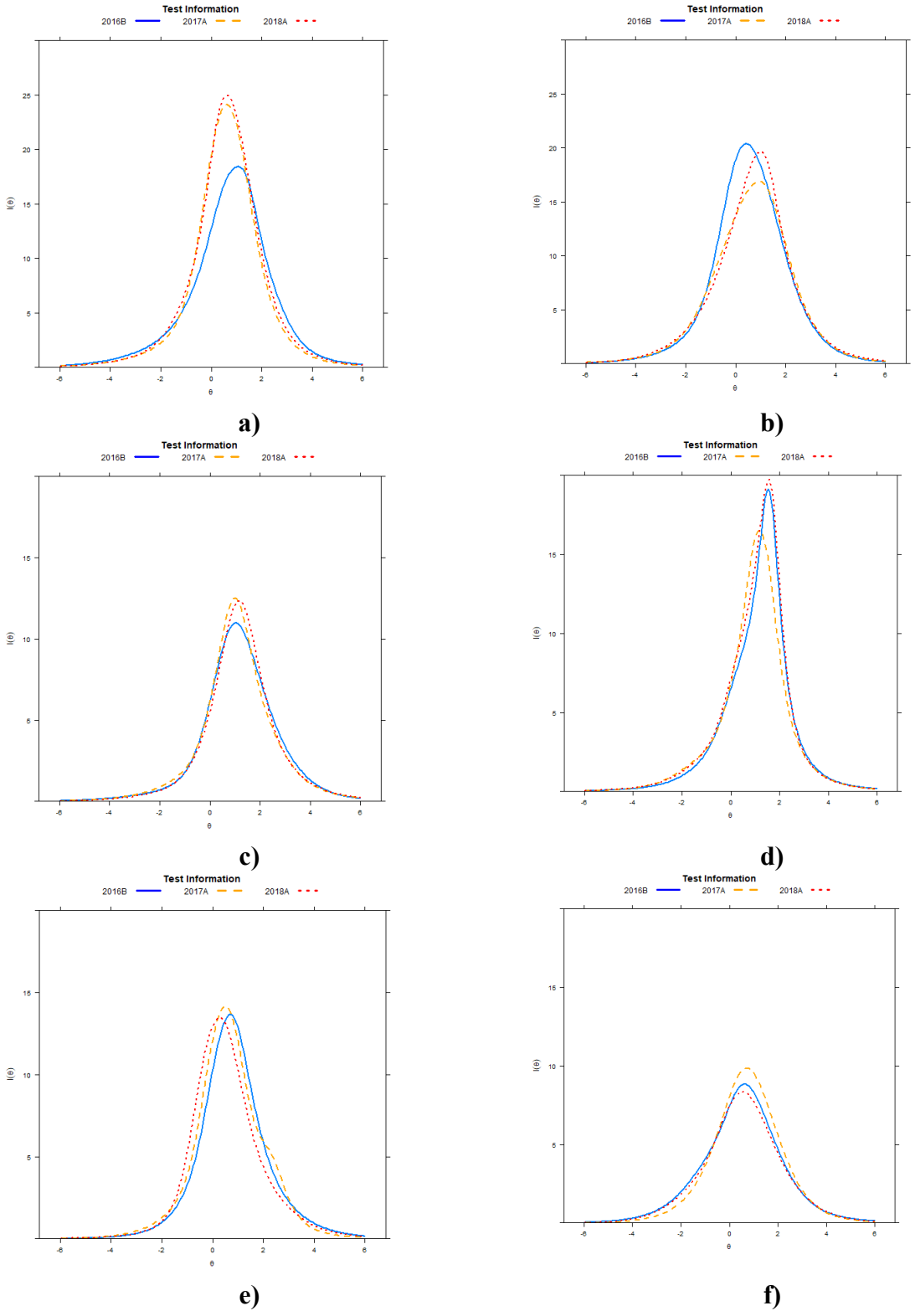


Figure 3. Equating transformations for the quant part in the left column (a and c) and verbal part in the right column (b and d) in Scenarios 1 (2018A->2016B, upper row) and 2 (2018A->2017A, lower row).

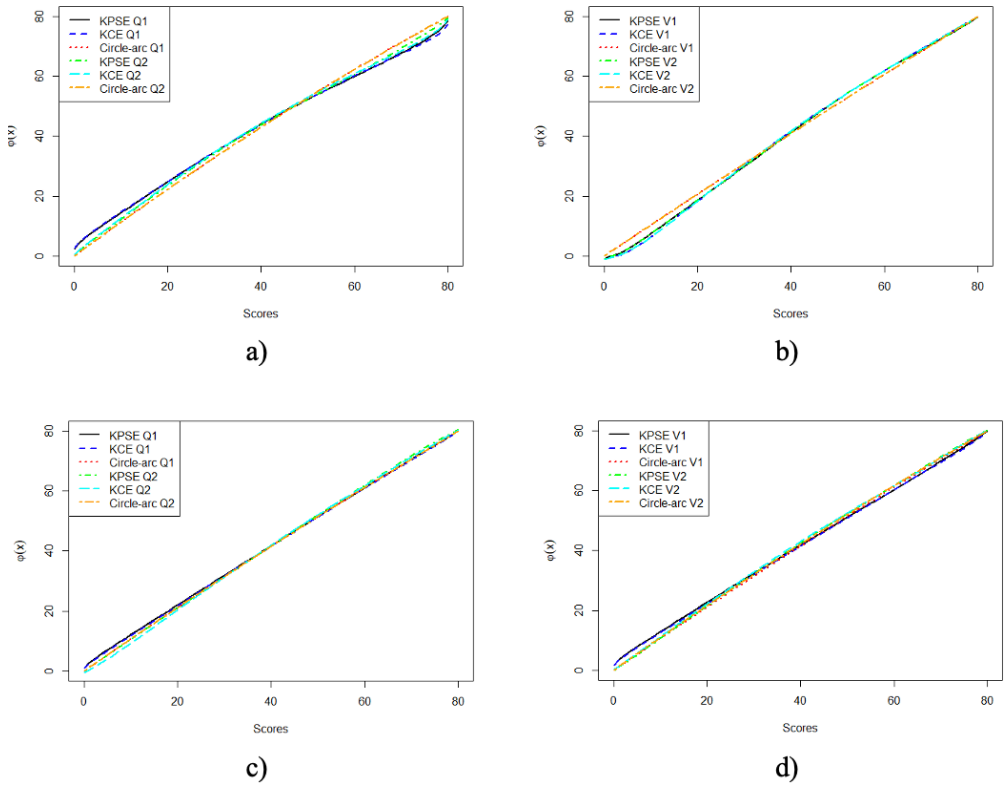


Figure 4. SEE values obtained in Scenarios 1 (upper row) and 2 (lower row) with the quantitative anchor test forms (right column, a and c) and verbal anchor test forms (left column, b and d).

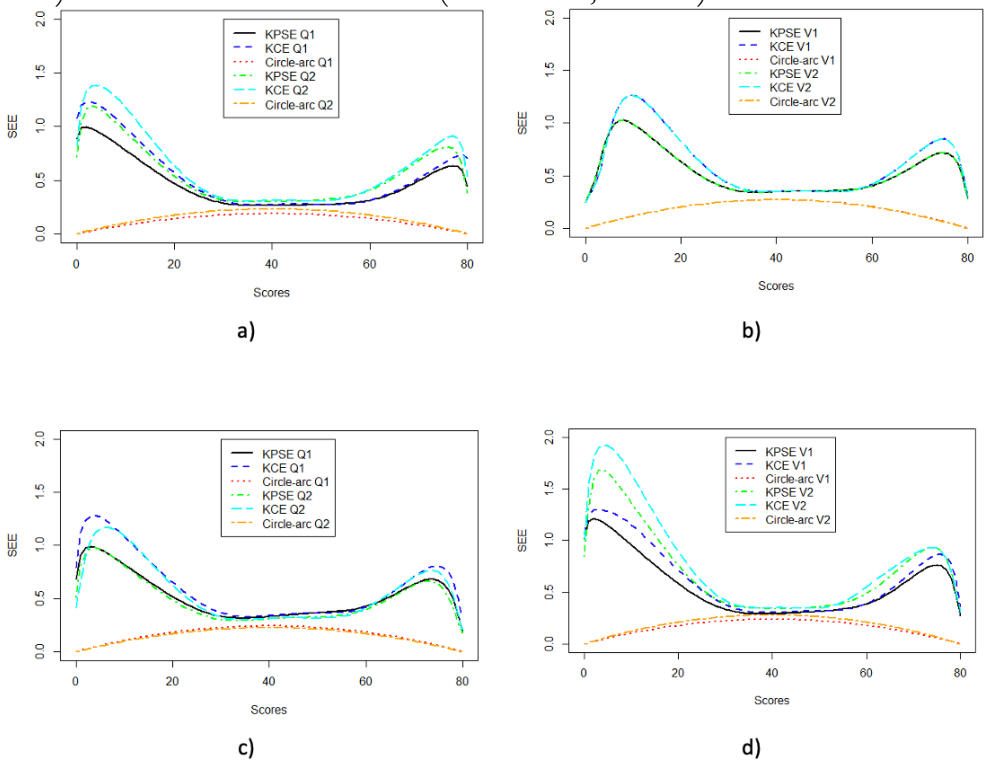


Table 3. The 23 scenarios (S) examined in the simulation study, together with the average correlations (C) from the simulations.

S	P=Q	P	Q	A_b	A_a	A_bS	A_aS	X_b	Y_b	C_{XA}	C_{YA}
S1	X									0.80	0.80
S2		+								0.80	0.80
S3a	X			+						0.76	0.77
S3b	X			-						0.80	0.79
S4a		+		+						0.79	0.79
S4b		+		-						0.80	0.79
S5a	X					+				0.79	0.78
S5b		+				+				0.79	0.79
S5c	X					-				0.79	0.80
S5d		+				-				0.81	0.80
S6	X				+					0.81	0.81
S7		+			+					0.80	0.80
S8a	X						+			0.80	0.80
S8b		+					+			0.80	0.80
S9a	X				-					0.80	0.80
S9b		+			-					0.80	0.80
S10a		+	+							0.80	0.80
S10b		-	-							0.77	0.78
S10c		-	+							0.77	0.80
S11a	X							+	+	0.78	0.78
S11b	X							-	-	0.79	0.80
S11c	X			+				+	+	0.77	0.78
S11d	X			-				-	-	0.80	0.80

P=Q: P and Q have similar ability, P = Group P is more (+) or less (-) capable. Q = Group Q is more (+) or less (-) capable, A_b = Anchor test form is more (+) or less (-) difficult. A_a = Anchor test form is more (+) or less (-) discriminating. A_bS = Anchor item difficulty is more (+) or less (-) spread. A_aS = Anchor item discrimination is more (+) or less (-) spread. X_b = Regular test form X is more (+) or less (-) difficult. Y_b = Regular test form Y is more (+) or less (-) difficult. C_{XA} (C_{YA}) = Average correlation of the simulated data with test form X (or Y) and anchor test form A in P (or Q).

normal distribution of ability. In the baseline case, we used the following item parameters: a~ item difficulty and item discrimination, by adding, subtracting, or multiplying selected constants to the item difficulty and item discrimination, by adding, subtracting, or multiplying selected constants to the item parameters, a method that has been successfully used in previous assessments of equating transformations under different conditions in simulations (e.g., Wiberg & van der Linden, 2011; van der Linden & Wiberg, 2010). We assessed effects of the test forms being more or less difficult by the addition (+) or subtraction (-) of 0.5 to the mean of item difficulty in the regular test forms (denoted X_b and Y_b) or anchor test form (A_b). Similarly, we examined effects of the test forms being more or less discriminating by multiplying the mean of item discrimination of the regular test forms (denoted X_a and Y_a) or anchor test form (A_a) by 1.5 (more) or 0.5 (less). The spread in the anchor item difficulty (A_bS) and discrimination (A_aS) were examined by multiplying the standard deviations of anchor item difficulty and discrimination by 1.5 (more spread) or 0.5 (less spread). The simulation procedure is summarized in Algorithm 1.

Algorithm 1.

1. Generate item parameters (a, b, c) for the regular and anchor test forms.
2. For 500 replications repeat the following:
   a. Generate abilities, θ, with the mvrnorm() function from the MASS R package.
   b. Generate scores for regular and anchor tests for P and Q groups with the three-parameter logistic

IRT model using the rmvlogis() function from the ltm R package.

c. Calculate sum scores for the regular and anchor test forms.

d. Perform the equating.

The correlations between the regular and anchor test forms were similar in all the considered scenarios, varying between 0.78 and 0.82, comparable to those observed in the real empirical data. The average correlation between the anchor and regular test forms in the 23 different scenarios are shown in the two last columns of Table 3 (C_{XA} and C_{YA}).

Equating methods

We used the three previously described equating methods: KCE, KPSE and Circle-arc. For the kernel equating methods we used a quadratic (second-order) polynomial model with one interaction term as a presmoothing model for the NEAT design, as we did in the empirical study. We are aware that it is better to try different models and use the best fitting model, but this approach was chosen here to limit the examined conditions. For the KPSE method, we set the weight to 0.5 as the sample sizes were the same.

Evaluation methods

To evaluate the equating transformation under the varied conditions, we examined the SEE, bias and WAB values.

Results of the simulation study

In Figures 5, 6, and 8-9 the SEE values obtained in the simulations are shown to the left and bias values

Figure 5. SEE (left) and bias (right) for the baseline case (s1) and the conditions when the groups had similar ability and the difficulty parameter in the anchor test form was varied (s3a, s3b, s5a and s5c).

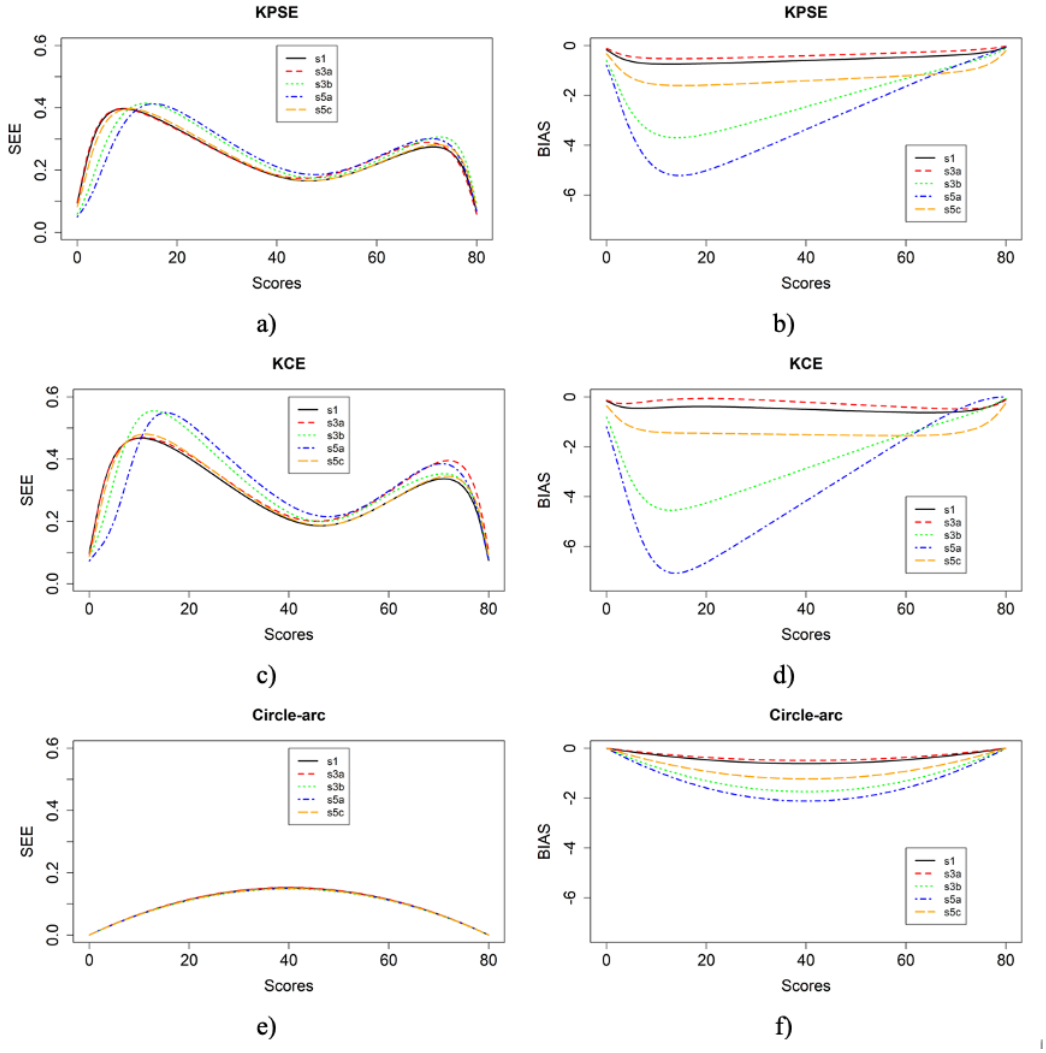
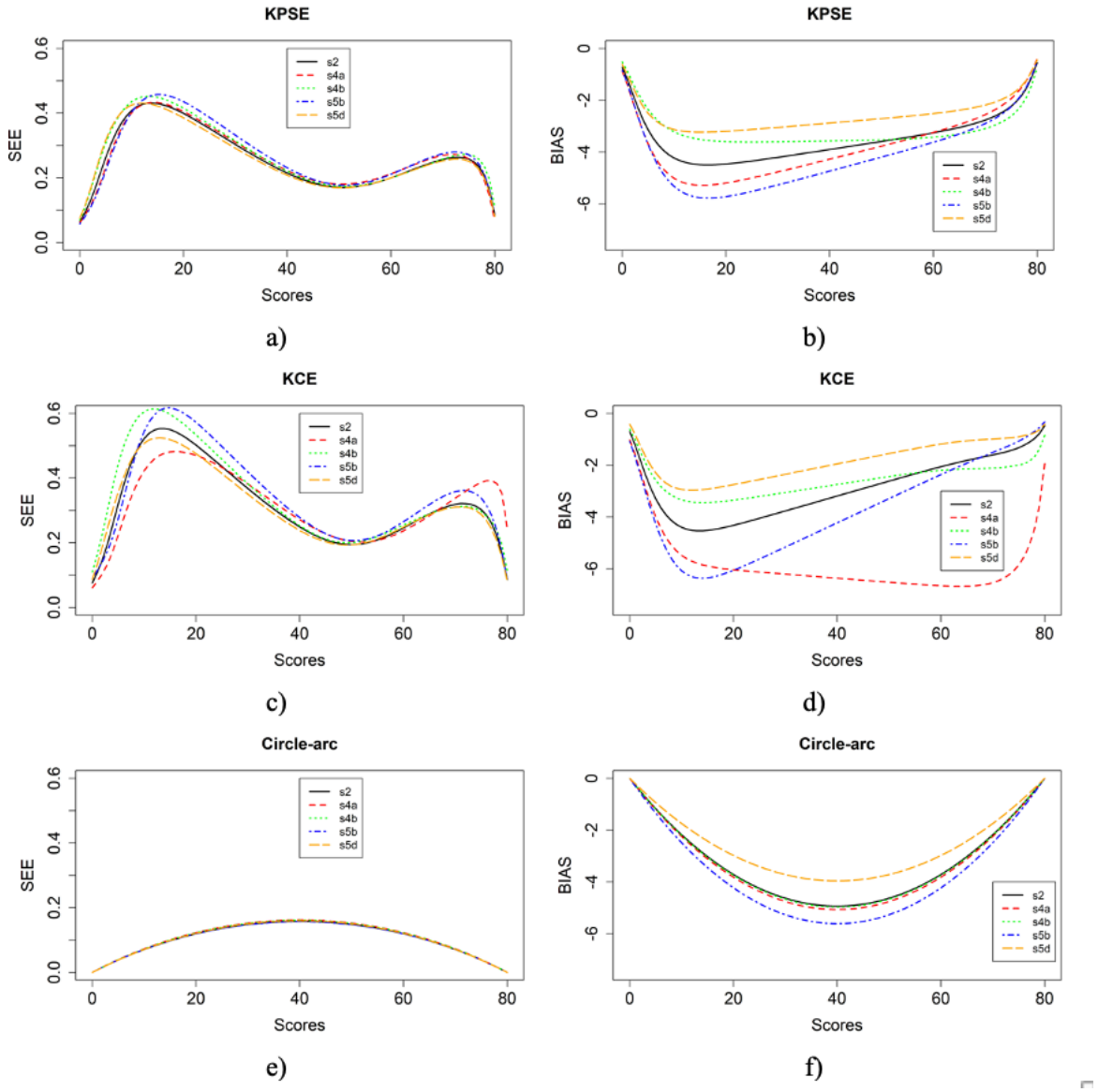


Figure 6. SEE (left) and bias (right) for groups differing in ability and difficulty of the anchor test form was varied (s2, s4a, s4b, s5b and s5d).



to the right. Figure 5 presents results of varying the difficulty parameter when the groups had similar ability levels. It shows that varying the difficulty parameter affected the equated scores (s3a, s3b, s5a and s5c), especially when using an easier anchor test form (s3b) or when the anchor test form difficulty is less spread in difficulty (s5c). The SEEs differed very little (with KPSE and KCE) or were almost identical (with Circle-arc equating) under the s1, s3a and s5c conditions, but differed somewhat more, especially with KCE, when the anchor test form was easier or more spread in difficulty than the regular test form (s3b and s5a). All the conditions involving a change in difficulty of the anchor test form affected the bias, except one scenario when the anchor test form was more difficult (s3a). Large between-condition differences were observed between scenarios when the anchor test form was easier than the regular test or anchor item difficulties were more spread (s3b, s5a) and the other scenarios (s3a and s5c) with both KPSE (Figure 5b) and KCE (Figure 5d). Corresponding differences were smaller with Circle-arc equating (Figure 5f).

WAB results presented in Table 4 clearly indicate that when groups had similar ability the Circle-arc method yielded the lowest weighted absolute bias compared to the other two studied equating methods

under most of the conditions. In the five scenarios presented in Figure 5, Circle-arc had the lowest WAB values in the three scenarios: s3b, s5a, and s5c, and KCE has the lowest WAB values in scenario s3a. In scenario s1, Circle-arc and KCE had similar WAB. WAB values in scenario s3a (the anchor test form was more difficult than the regular test form) were the lowest compared to the rest of the studied conditions for the three methods.

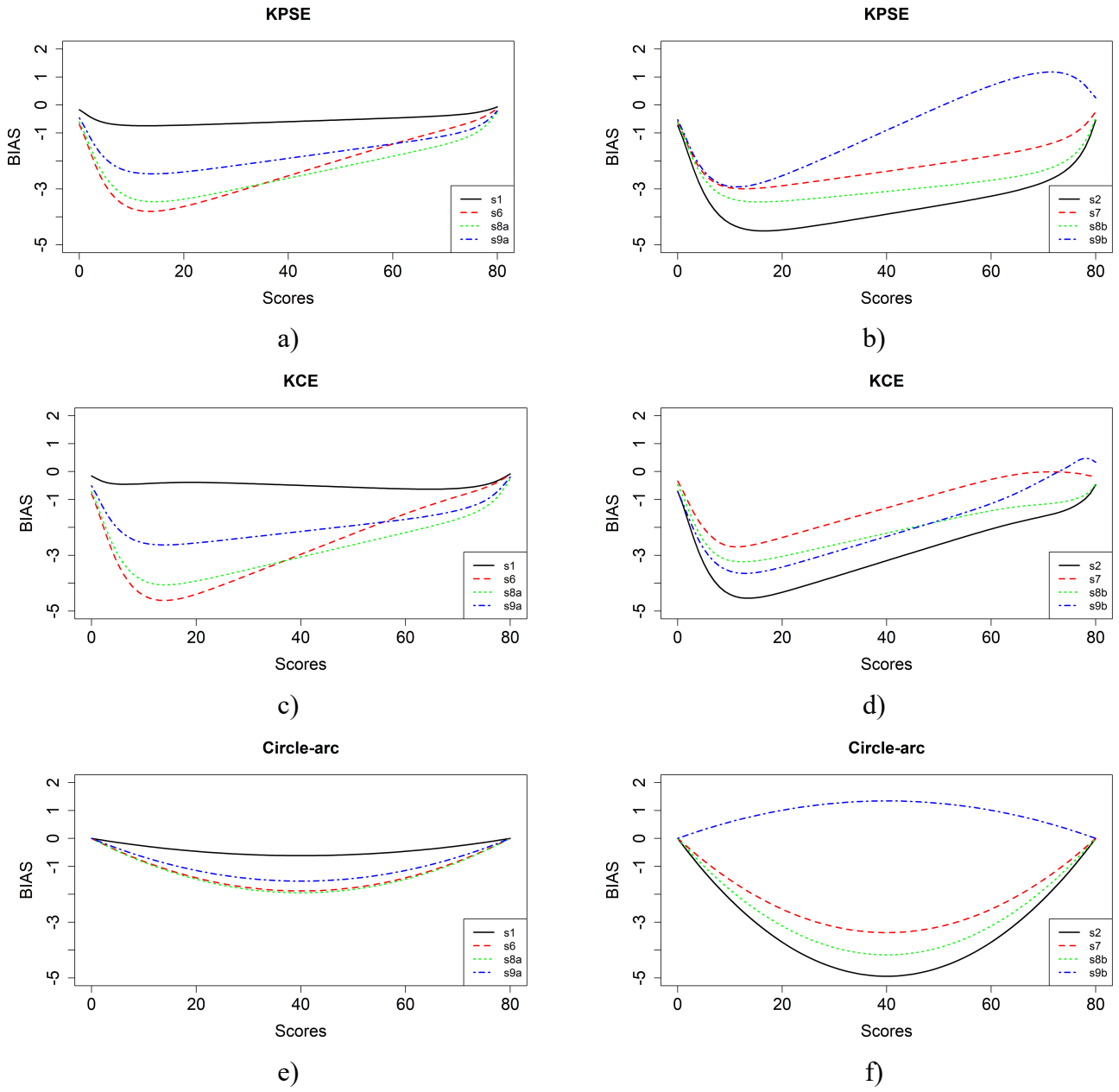
Figure 6 displays results corresponding to those depicted in Figure 5. However, in this case, scenarios involve one group with average ability and another with higher ability, as opposed to both groups possessing the same level of ability. The differences in SEE values were small under all five considered scenarios with both KPSE and Circle-arc equating methods (Figures 6a and 6e) and vary more with KCE (Figure 6c). The changes in the bias values, were largest when the anchor test form was more difficult than the regular test form (s4a) with KCE and when the anchor test form was more spread in difficulty (s5b) with both KCE and Circle-arc equating, both when the groups had similar ability (Figure 5) and differing ability (Figure 6). The WAB results (Table 4) indicate that when the groups had differing ability the KCE method yielded the lowest weighted absolute bias compared to the other two studied equating methods under most of the conditions. WAB was largest under scenario s5b (when the anchor test form was more spread in difficulty than the regular test form) with both KPSE and Circle-arc equating methods, and under s4a (when the anchor test form was more difficult than the regular test form) with KCE method.

Figure 7 shows the bias values when varying the discrimination parameter, both when groups had similar ability (s1, s6 and s8a) and different abilities (s2, s7 and s8b). The size and spread of the item discrimination parameter substantially affected the bias, both when groups had similar ability (left panels)

Table 4. Descriptive statistics of weighted absolute bias (WAB) over 500 replications under all studied conditions with the three equating methods.

	WAB		
	KPSE	**KCE**	**Circle-arc**
Scenario	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**
s1	0.57 (0.16)	**0.53 (0.18)**	0.54 (0.13)
s2	3.52 (0.17)	**2.60 (0.19)**	4.08 (0.13)
s3a	0.39 (0.16)	**0.32 (0.15)**	0.43 (0.13)
s3b	2.20 (0.17)	2.56 (0.20)	**1.54 (0.13)**
s4a	**3.69 (0.17)**	6.47 (0.20)	4.19 (0.13)
s4b	3.45 (0.17)	**2.47 (0.19)**	4.11 (0.13)
s5a	2.97 (0.19)	3.61 (0.22)	**1.87 (0.13)**
s5b	4.11 (0.17)	**3.24 (0.20)**	4.64 (0.13)
s5c	1.36 (0.17)	1.51 (0.19)	**1.09 (0.13)**
s5d	2.66 (0.17)	**1.57 (0.19)**	3.27 (0.13)
s6	2.27 (0.17)	2.64 (0.19)	**1.66 (0.13)**
s7	2.06 (0.17)	**0.83 (0.17)**	2.79 (0.13)
s8a	2.43 (0.17)	2.86 (0.20)	**1.72 (0.13)**
s8b	2.85 (0.16)	**1.80 (0.19)**	3.45 (0.13)
s9a	1.78 (0.17)	2.04 (0.20)	**1.35 (0.13)**
s9b	**0.96 (0.16)**	2.04 (0.23)	1.18 (0.13)
s10a	2.02 (0.18)	2.37 (0.20)	**1.34 (0.13)**
s10b	2.27 (0.17)	2.68 (0.20)	**1.70 (0.14)**
s10c	**1.07 (0.16)**	2.88 (0.29)	3.60 (0.14)
s11a	2.25 (0.17)	2.65 (0.20)	**1.71 (0.14)**
s11b	2.12 (0.18)	2.53 (0.21)	**1.43 (0.13)**
s11c	2.27 (0.17)	2.68 (0.20)	**1.70 (0.14)**
s11d	2.02 (0.18)	2.37 (0.20)	**1.34 (0.13)**

Figure 7. Bias values for KPSE (a,b), KCE (c,d), and Circle-arc equating (e,f) when varying the discrimination parameter with groups similar in ability to the left (s1, s6, s8a, and s9a) and groups with differing abilities (s2, s7, s8b, and s9b).



and different abilities (right panels). The bias was largest under scenario s2 with groups differing in ability, using all three equating methods. When groups had similar abilities, the bias values obtained with all three methods were largest when the anchor test form was less discriminating than the regular test form (s6) or anchor item discrimination was more spread (s8a). The SEE values were similar in all these scenarios, so the figures are excluded but can be provided upon request from the corresponding author.

Similarly to the results presented in Figure 7, the WAB values were largest (Table 4) in scenario s2 for KPSE and Circle-arc, and for scenario s8a for KCE. When groups had differing abilities, KCE produced the lowest bias compared to the other two methods (conditions presented in right panels in Figure 8),

except in scenarios9b where KPSE yielded the lowest bias.

Figure 8 shows that the SEE varied substantially when the groups had different abilities. Only th Circle-arc exhibited negligible differences in the SEE for all different combinations of group abilities. For KPSE and KCE, the SEE was high for all test scores when the ability differences were large (s10c). Furthermore, the SEE was large for low test scores and small for higher test scores when both P and Q groups had higher than average abilities (s10a). Similarly, the SEE was small for low test scores and large for high test scores when the groups had lower than average abilities (s10b). KCE generally yielded slightly higher SEE values than KPSE in all scenarios with ability differences. Differences in the bias were largest when either one group had higher abilities (s2), or one group had a low ability level and the other a high ability level (s10c). Note that the Circle-arc method yielded large

Figure 8. SEE (left) and bias (right) when we there are differences in abilities of the groups.

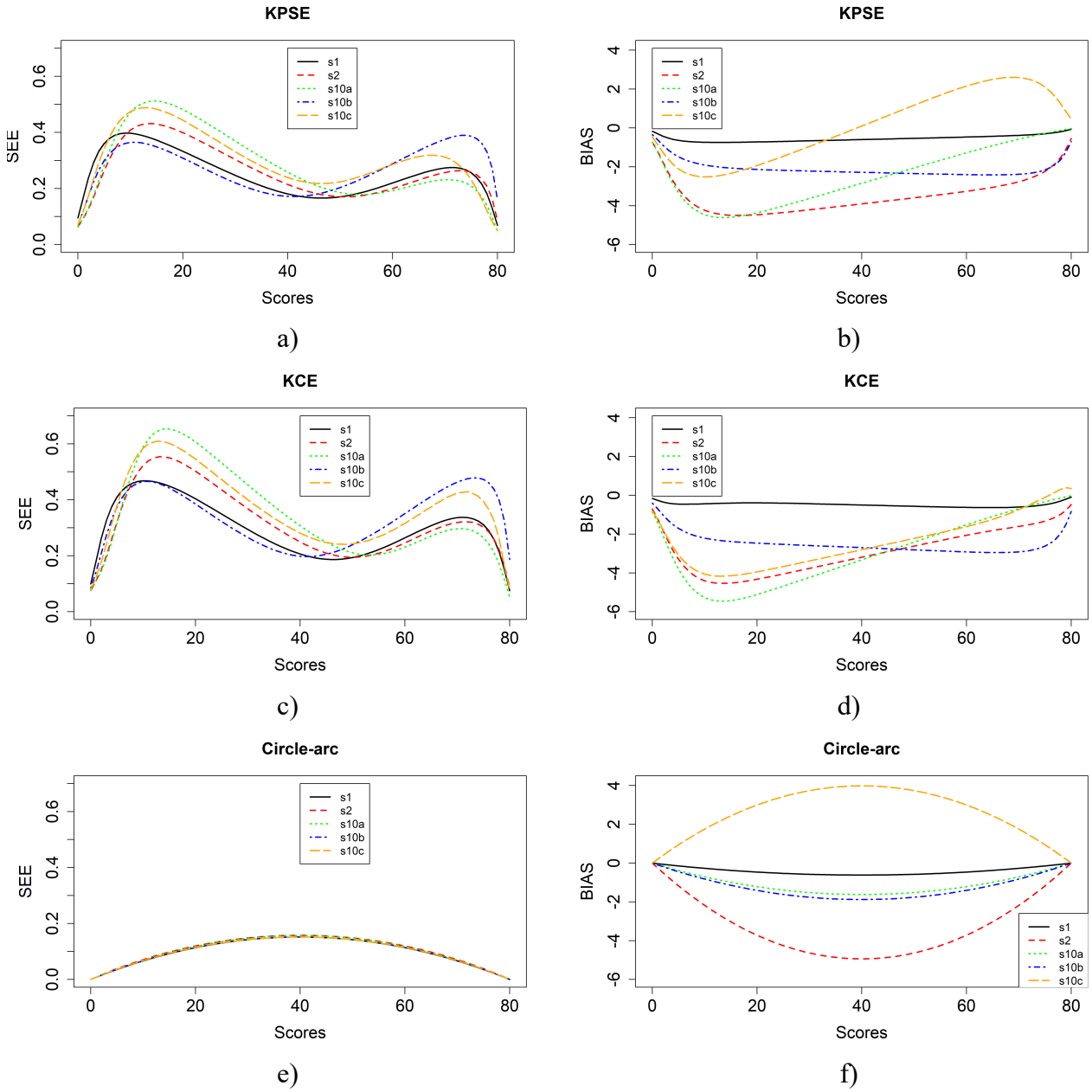
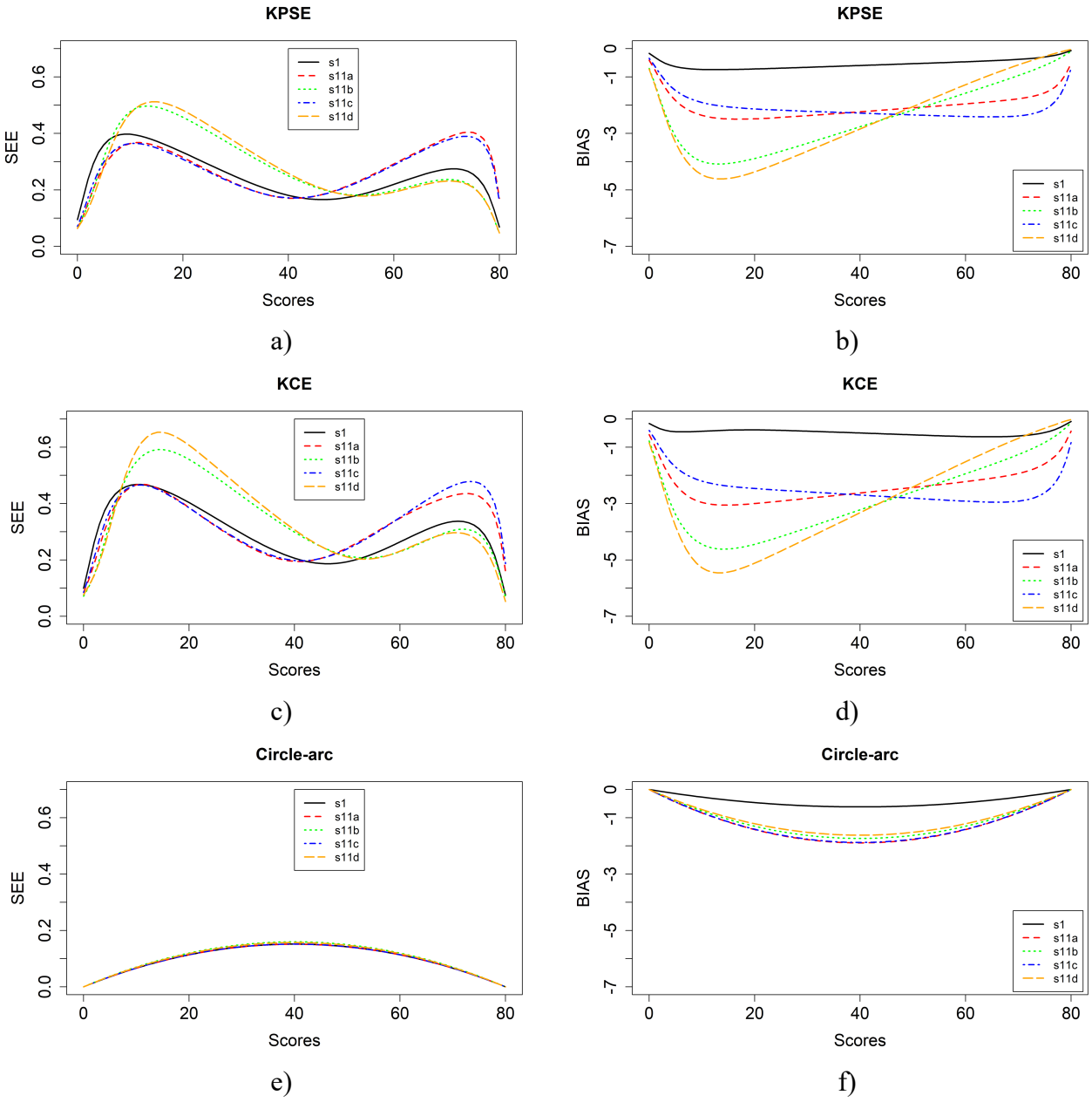


Figure 9. SEE (left) and bias (right) when varying the difficulty of the regular test forms (s11a and s11b) and anchor test form (s11c and s11d) when groups have similar ability.



bias and WAB values than the kernel equating methods when there were differences in abilities between the groups (s2 and s10c).

Finally, we examined the effects of varying the difficulty of the regular test forms for groups of similar abilities on equating results (Figure 9). Increasing the difficulty of the regular test form had only small effects on results obtained using all three equating methods, with both an unchanged anchor test form and more difficult anchor test form (s11a and s11c). Similarly, decreasing difficulty of the test form had small effects on results obtained using all three equating methods, with both an unchanged anchor test form and less difficult anchor test form (s11b and s11d). However, there was a large difference in equating results between these two sets of conditions. When using an easier regular test form and an easier anchor test form the SEE values obtained with KCE and KPSE were inversely related to the test scores, and an opposite

pattern was observed when using a more difficult regular test form and a more difficult anchor test form. The SEE was low and similar under all these scenarios when using Circle-arc. The differences in the bias values were largest for lower test scores when easier regular and/or anchor test forms were used (s11b and s11d) with kernel equating methods. The opposite pattern was not seen when test forms (regular or anchor) were more difficult (s11a and s11c). In terms of WAB (see Table 4), highest average values were obtained with KCE (2.37 – 2.68), followed by KPSE (2.02 – 2.27) and Circle-arc equating (1.34 – 1.71).

Discussion

The overall aim of this study was to examine effects of differences in group ability and features of the anchor test form on equating transformations and SEE values using real college admissions test (SweSAT) data and simulations with variations in test form features (which also addressed bias). The empirical study indicated that results of the three considered equating methods only slightly differed. Differences between them when using different anchor test forms were smaller than differences between the scenarios (with test-taking groups of differing abilities), especially for the verbal test forms. KPSE and KCE yielded similar SEE patterns, with higher values in the lower and upper score ranges than in the mid score ranges. However, KPSE consistently yielded lower SEE values than KCE.

Circle-arc equating yielded the lowest SEE values among the three examined methods in both scenarios considered in the empirical study, and at all test scores. Circle-arc equating has previously been typically recommended when samples are small, e.g., by Livingston & Kim (2009), but it was included here as it is one of the methods used to equate SweSAT forms in practice. The generally low SEE values are probably due to the definition of SEE for Circle-arc equating. The method only estimates the mean, a high point, and a low point, which is why the equating is always done in only one direction. This could be problematic in practice if one of the test forms is more difficult for some score points. To examine differences in anchor test groups and anchor test forms more thoroughly, and enable generalization of our findings, we performed a simulation study, designed to mirror real testing conditions by building on the observed test score distributions but have more generality than if we had just sampled SweSAT items.

The simulation study included 23 scenarios, obtained by varying the ability of the groups that received the anchor test forms, the item difficulty and item discrimination of the anchor test forms, and the item difficulty of the regular test forms. Previous studies have mainly focused on the anchor test form's features (e.g., Sinharay, 2018; Sinharay & Holland 2006a; 2006b and Trierweiler et al., 2016) and not on the groups given it, or the interactions between its features and differences in the groups. The findings that item difficulty and the abilities of test-taking groups influence both bias and SEEs are in line with previous research (see e.g., Cook & Petersen, 1987; Kolen, 1990; Kolen & Brennan, 2014; Sinharay & Holland, 2007; Liu et al., 2011; Gonzalez & Wiberg, 2017). Changes in item discrimination resulted in variation of bias values but did not have any large impact on the SEE. This is not surprising as similar results were obtained by van der Linden and Wiberg (2010) for equated values in local equating and Wiberg and Gonzaléz (2021) for equated values of mixed-format tests. Circle-arc equating yielded the lowest WAB values compared to the kernel equating methods when groups had similar abilities, while KCE yielded the lowest WAB values when groups had differing abilities.

Equating with an easier anchor test form and/or easier regular test form, as well as an anchor test with more spread difficulties, resulted in higher SEE and higher bias values compared with using average or more difficult anchor test forms, especially at the lower test scores. This is probably because both capable and less capable test takers managed to get the easier items correct than when test forms of average difficulty were used. The KCE method was more sensitive to item difficulties when groups had similar abilities, and to groups' abilities when both groups were less able or both groups were more able. Circle-arc equating, on the other hand, was more sensitive to item difficulties when groups had differing abilities. The reason for this behavior is probably due to the few data points used when calculating it. KPSE had the lowest WAB values, compared to KCE and Circle-arc, when one group was less capable ($\theta \sim N(-0.5,1)$) and the other group more capable ($\theta \sim N(0.5,1)$ than average. This is in line with Powers and Kolen (2014) who found that chained equating was less sensitive to group differences

(although they didn't study extreme differences as we did) than the frequency estimation method, but they did not include the KCE and KPSE kernel methods or the Circle-arc method in their study. Liu et al. (2011) compared chained and poststratification equating using real data and found that increases in groups' abilities resulted in larger bias when using a criterion equating function, regardless of equating method. Our study confirmed this, although we used kernel equating methods. Moreover, our results also indicate that bias is larger when both equated groups have similar low abilities than when groups have similar high abilities.

The obtained results can be used in several ways in practice. First, equating results, standard error of equating and bias, are optimal when the anchor and regular test forms are of average difficulty, which in our case was when b~N(0.4,1). The results concerning item difficulty of the anchor test forms are in line with previous research (e.g., Sinharay, 2018; Sinharay & Holland, 2006a; 2006b). However, our study provides further illumination, as previous studies did not examine effects of variations in either ability levels in this context or difficulty levels of regular test forms in combination with variations in anchor test form difficulty. From our results we cannot recommend using an easy anchor test form together with easy regular test forms, or an anchor test form with a wider spread of difficulties. Second, if the CTT framework (and thus a non-IRT equating method) is being used and there is an option to choose which test-taking group will receive the anchor test form, it should be given to a group of average ability if possible. Giving it to a high ability group will have a negative effect on the equating in terms of the size of the SEE, especially for low test scores. If it is given to a low ability group there will be a negative effect in terms of SEE for the higher test scores. The bias will also be larger in both these cases. This is especially important if an anchor test form is only being administered at a limited number of test centers and the ability level of test-takers is known to differ between the centers, as it is for the SweSAT used in the empirical study. Third, KPSE yielded lower SEE values but larger bias than KCE in most of the scenarios. Circle-arc yielded the lowest SEE values in all the studied scenarios, but it resulted in the largest bias values when there were differences in the groups' abilities. Thus, if it is known that there are large differences in the groups' abilities, we recommend using KCE, but Circle-arc equating could be used if there are no or only small differences in their abilities. Note, Circle-arc equating seems to be possible to use with larger sample sizes, even though it was initially designed for use with small sample sizes.

This study has several limitations. First, we did not examine effects of differences in length of the anchor test forms, but this was partly because they have been previously addressed (e.g., Sinharay & Holland, 2006b; Ricker & von Davier, 2007). Another reason is that we wanted to make it difficult for test takers to identify the anchor test form, so it had to be the same length as all the regular parts of the SweSAT, i.e., half the length of the quantitative or verbal section, each of which is given as two equal parts. Although effects of varying its length could have been assessed in the simulation study, our overall aim was to identify optimal features of both an anchor test form that cannot be easily detected by the test takers and groups to receive it. Second, we only considered CTT equating methods because the SweSAT is rooted in CTT. In the future it would be interesting to repeat this study with IRT equating methods. In that case we do not expect to see any impact of group differences on the equated values as IRT methods are generally sample-independent, in contrast to CTT methods which have known sample-dependence. Note, some readers may question our simulation of data using a unidimensional IRT model, as a multidimensional IRT model may have reflected the reality better. However, this is not a problem with SweSAT datasets as a dimensionality analysis by Wedman and Lyrén (2019) concluded that they have one verbal dimension and one quantitative dimension. Thus, as we examined the verbal and quantitative parts of the test forms separately, we could use a unidimensional IRT model. Note, none of the three CTT methods performed well in some of the scenarios, e.g. when anchor item difficulty was more spread (s5b) or when anchor item was more discriminating (s6). Thus, in the future it would be interesting to assess the performance of other equating methods in the scenarios where these methods did not perform well. Third, we chose to use linear equating as the criterion function in this study when calculating the bias. Different choices of a criterion function can potentially lead to different results and thus this is a topic that should be examined more in the future. Finally, Wiberg (2021) found that equating transformations are influenced by the linkage plan. Thus, more comprehensive examination of the impact

of various anchor test forms when using different linkage plans would also be interesting, as well as linkage plan effects on the optimal anchor test form and optimal test group to receive it.

To summarize, when using the three examined methods the features of the anchor test form clearly affect the equating transformation. It is also very important to decide which group receives the anchor test form as abilities of that group can substantially affect the equating transformation.

References

Albano, A. D. (2016). Equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8),1–36. https://doi.org/10.18637/jss.v074.i08

Andersson, B., Bränberg, K. & Wiberg, M. (2013). Performing the kernel method of test equating using the package kequate. *Journal of Statistical Software, 55*, 1–25. https://doi.org/10.18637/jss.v055.i06

Angoff, W. H. (1968). How we calibrate College Board scores. *College Board Review, 68,* 11–14.

Chalmers RP (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244. https://doi.org/10.1177/014662168701100302

Divgi, D. R. (1987). *A stable curvilinear alternative to linear equating* (Rep. No. CRC 571). Alexandria, VA: Center for Naval Analyses.

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating* (ETS SR-98-02). Princeton, NJ: ETS.

González, J. & Wiberg, M. (2017). *Applying test equating methods using R.* Cham, Switzerland: Springer. https://doi:10.1007/978-3-319-51824-4

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education/Praeger.

Häggström, J. & Wiberg, M. (2014). Optimal bandwidth in observed-score kernel equating. *Journal of Educational Measurement, 51*(2), 201–211. https://doi.org/10.1111/jedm.12042

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, *22*(3), 197–206. https://doi.org/10.1111/j.1745-3984.1985.tb01058.x

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, 3, 97–104. https://doi.org/10.1207/s15324818ame0301_7

Kolen, M. J., Brennan, R. L., & Kolen, M. J. (2014). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Liu, J., Sinharay, S., Holland, P., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT data. *Journal of Educational Measurement, 48*, 361–379. https://doi.org/10.1111/j.1745-3984.2011.00150.x

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2011). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement, 71*, 346–361. https://doi.org/10.1177/0013164410375571

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, *3*, 73–95. https://doi.org/10.1207/s15324818ame0301_6

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46, 330–343. https://doi.org/10.1111/j.1745-3984.2009.00084.x

Lyrén, P.-E. (2009). *A perfect score: Validity arguments for college admission tests* (PhD dissertation, Institutionen för Beteendevetenskapliga mätningar, Umeå Universitet). Retrieved from

https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-25433

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. *Test equating*, 71–135. New York: Academic Press.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education.

Powers, S. & Kolen, M. J. (2014). Evaluating equating accuracy and assumptions for groups that differ in performance. *Journal of Educational Measurement, 51*(1), 39-56. https://www.jstor.org/stable/24018322

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54-75. https://doi.org/10.1111/j.1745-3984.2009.00099.x

Ricker, K. L., & von Davier, A. A. (2007). The impact of anchor test length on equating results in a nonequivalent groups design. *ETS Research Report Series*, *2007*(2), i-19. https://doi.org/10.1002/j.2333-8504.2007.tb02086.x

Sinharay, S. (2018). On the choice of anchor test in equating. *Educational Measurement: Issues and Practice*, *37*(2), 64–69. https://doi.org/10.1111/emip.12175

Sinharay, S., & Holland, P. W. (2006a). *The correlation between the scores of a test and an anchor test* (ETS RR-06-04). Princeton, NJ: Educational Testing Service.

Sinharay, S. & Holland, P. (2006b). Choice of anchor test in equating. ETS report RR-06-35.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*(3), 249–275. https://doi.org/10.1111/j.1745-3984.2007.00037.x

Sinharay, S., & Holland, P.W. (2010a). The missing data assumptions of the NEAT design and their implications for test equating. *Psychometrika*, *75*, 309–327. https://doi.org/10.1007/s11336-010-9156-6

Sinharay, S., & Holland, P.W. (2010b). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, *47*, 261–285. https://doi.org/10.1111/j.1745-3984.2010.00113.x

Sinharay, S., Haberman, S., Holland, P., & Lewis, C. (2012). A note on the choice of an anchor test in equating. *ETS Research Report Series*, *2012*(2), i-9. https://doi.org/10.1002/j.2333-8504.2012.tb02296.x

Sinharay, S., Holland, P. W., & von Davier, A. A. (2011). Evaluating the missing data assumptions of the chain and poststratification equating methods. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 281–296). New York: Springer.

Trierweiler, T.J., Lewis, C, & Smith, R.L. (2016). Further study of the choice of anchor tests in equating. *Journal of Educational Measurement, 53*(4), 498–518. https://www.jstor.org/stable/45148405

van der Linden, W. J. & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement. 34*(8), 620–640. https://doi.org/10.1177/0146621609349803

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The kernel method of equating. New York: Springer.

von Davier, A. A., Holland, P.W., & Thayer, D. T. (2004b). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, *41*, 15–32. https://doi.org/10.1111/j.1745-3984.2004.tb01156.x

Wallin, G., Häggström & Wiberg, M. (2021). How important is the choice of bandwidth in kernel equating? *Applied Psychological Measurement, 45*(7-8), 518–535. https://doi.org/10.1177/01466216211040

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design.

Applied Psychological Measurement, *32*, 632–651. https://doi.org/10.1177/01466216083149

Wang, S., Zhang M., & You S. (2020). A comparison of IRT observed score kernel equating and several equating methods. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.00308

Wedman, J., & Lyrén, P.-E. (2015). Methods for examining the psychometric quality of subscores : A review and application. *Practical Assessment, Research & Evaluation*, 20. Published. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-112181

Wiberg, M. (2021). On the use of different linkage plans with different observed-score equipercentile equating methods. *Practical Assessment, Research and Evaluation,* *26*(24) 1–16. https://doi.org/10.7275/21481778

Wiberg, M. & González, J. (2017). Statistical assessment of estimated transformations in observed-score equating, *Journal of Educational Measurement*, *53*(1), 106–125. https://doi.org/10.1111/jedm.12103

Wiberg, M. & González, J. (2021). *Possible factors which may impact kernel equating of mixed format tests.* In Wiberg, M., Molenaar, D., González, J., Böckenholt, U., & Kim, S-J. (Eds.). Quantitative Psychology – 85th Annual Meeting of the Psychometric Society, Virtual, 2020, Cham: Springer. 199-206.

Wiberg, M. & van der Linden, W. J. (2011). Linear local observed-score equating. *Journal of Educational Measurement,* *48*(3), 229–254. https://doi.org/10.1111/j.1745-3984.2011.00148.x

Citation:
Laukaityte, I., & Wiberg, M. (2024). Impacts of differences in group abilities and anchor test features on three Non-IRT test equating methods. *Practical Assessment, Research, & Evaluation*, 29(5). Available online: https://doi.org/10.7275/pare.2020

Corresponding Author:
Inga Laukaityte
Umeå University

Email: inga.laukaityte [at] umu.se