Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 23, November 2021 ISSN 1531-7714

On the Use of Different Linkage Plans with Different Observed-score Equipercentile Equating Methods[1]

Marie Wiberg, *Umeå University*

The overall aim was to examine the equated values when using different linkage plans and different observed-score equipercentile equating methods with the equivalent groups (EG) design and the nonequivalent groups with anchor test (NEAT) design. Both real data from a college admissions test and simulated data were used with frequency estimation, chained equating, and kernel equating methods. The overall results were that different linkage plans gave different equated values and standard errors (SEs) both in the EG design and in the NEAT design, and there were some differences between the equating methods. The simulation study confirmed the empirical results and suggested that the kernel equating methods gave lower SEs in the examined conditions and had fewer differences that mattered compared with traditional equating methods when different linking plans were used. Finally, when one of the test forms was easier or if the ability of the test groups differed, the choice of linkage plan had more of an impact than when the test forms or test groups were more similar.

Introduction

It is important to examine the quality of a test over time when a standardized achievement test is administered consecutively with different test forms over several years (Wiberg & von Davier, 2017). The main focus here is on how the choice of a linkage plan affects the equating of test scores, i.e. the statistical models and methods that are used to make test scores comparable among different test forms, so that the scores can be used interchangeably (González & Wiberg, 2017). The overall aim was to compare equated values and standard errors when using traditional observed-score equipercentile equating methods (Kolen & Brennan, 2014) and kernel equating methods (von Davier et al., 2004a) with different linkage plans with both real college admissions test data and simulated test data. As different data collection designs may give different results (see e.g., Wiberg & Bränberg, 2015 or Wallin & Wiberg, 2019), both an equivalent groups (EG) design and a nonequivalent group with anchor test (NEAT) design were examined.

A number of research studies have focused on the comparison of different equating methods in general. Livingston et al. (1990) studied which combinations of sampling and equating methods work best by comparing Tucker, Levine equally reliable, chained equipercentile, frequency estimation, and IRT equating methods with the three-parameter logistic model with either representative samples or matched samples. They

[1] The research was funded by the Swedish Wallenberg MMW 2019.0129 grant.

concluded that the IRT and Levine methods agreed with each other, and that the chained equipercentile method had low bias in the representative samples. Mao et al. (2006) examined real data and found only trivial result differences when comparing traditional equipercentile equating with kernel equating in the EG design and with post-stratification equating with a NEAT design. In Liu and Low (2008), the use of traditional and kernel equating methods were examined in two conditions: equating to a very different population and equating to a similar population. Their conclusions were that traditional and kernel equating methods were comparable and gave similar results when the populations were similar on the anchor score distribution even though they rely on different assumptions. Note, they also concluded from their studies that if the test group changed, the equating methods gave different results.

When equating several test forms over a number of administrations, it is possible to use several different linkage plans. Previous research on linkage plans has focused on the accumulated equating error of using several linkings between two test forms or different linking plans with empirical data from real tests (see e.g. Guo, 2010; Guo, Liu, Dorans, & Feigenbaum, 2011; Haberman, Guo, Liu, & Dorans, 2008; Taylor & Lee, 2010). For example, Guo (2010) used different linear equating methods and examined chains of equating within the NEAT design. Liu, Curley, and Low (2009) re-administered and re-equated an old test form in order to assess the amount of equating errors that occurred over time. Moses, Deng, and Zhang (2011) examined the effect of using more than one anchor test. Puhan (2009) used empirical test data from three tests with cut scores to compare the amount of equating errors for different linkage patterns within the NEAT design using chained linear and chained equipercentile equating, and within the EG design using equipercentile equating and linear equating. Livingston and Antal (2010) used a NEAT design to examine different linkage plans with empirical data and suggested how to adjust the final equating when several linkage plans are used. There has also been analytical research about the random and systematic errors of multiple equatings (Haberman & Dorans, 2011) and chains of equating (Haberman, 2010). A common conclusion from these studies is that using different linkage plans give different equated values; however, it is not completely clear how different factors affect the equated values.

Kolen and Brennan (2014) discussed different hypothetical linkage plans without using any data. They proposed the following four rules for linkage plans: 1) Avoid equating strains by minimizing the number of links that affect the comparison of scores on test forms given at successive times, 2) link to the same time of the year if possible, 3) minimize the number of links connecting each test form back to the initial one, and 4) avoid linking back to the same test form too often. They also suggested that one possibility to check a conducted equating is to equate the old test form back to the newest one, i.e. to perform a circular equating. Recently, Wiberg (2017) compared traditional equating methods with kernel equating methods and IRT equating methods when two different linkage plans were used with real test data. Her conclusions were that different equating methods and different linkage plans gave somewhat different results, but it was not clear under which conditions as only empirical data was used.

The present study differs from previous studies in a number of important aspects. First, to the best of our knowledge, previous studies were either analytical or administered a real test numerous times. In this study, we include both real college admissions test data and a simulation study where different conditions were examined. Second, in the empirical study with the real college admissions test data, we deliberately included a case where we violated rule 2, i.e. we compared test forms given at different time points in a year, as we have not seen that examined in previous studies. By examining different lengths of linking plans, we also examined rule 1 – the use of different lengths of equating strains. Third, previous studies have focused on traditional equating methods. In this study, we included both traditional observed-score equipercentile equating methods and compared them with different kernel equating methods within both the EG and the NEAT designs with respect to the different linkage plans. Fourth, we also examined different conditions, including whether one of the test groups was more capable or if one test form was easier. Fifth, we examined both the equated values and the standard errors – a combination which has not been examined for different linkage plans before.

The rest of the paper is structured as follows. In the next section, short descriptions of the used equating methods are given, followed by a description of the empirical study, and the results from the empirical study. Next, the set-up of the simulation study is presented, which is followed by the results from the simulation study. The paper ends with a discussion with some concluding remarks and suggestions for future studies.

Equating methods

We used observed-score equipercentile equating methods in our study and the methods used are described briefly below for the NEAT design, although we also used frequency estimation and kernel equating for the EG design, labelled FG and KG respectively, in the later result sections. For the NEAT design, we included frequency estimation, chained equipercentile equating, chained kernel equating, and post-stratification kernel equating. Both the empirical study and the simulation study were performed in R, and the equating methods were used as described in González and Wiberg (2017).

Frequency estimation

In frequency estimation (Angoff, 1971), we assume that for both test forms X and Y, the conditional distribution of the total score given each anchor score is the same in both populations. Thus, one can estimate the cumulative score distributions (CDF) of the test forms X and Y in populations P and Q, respectively, for a target population T when a common anchor test A is administered. Let *x* be the test scores on test form X, let *y* be the test scores on test form Y, and let *a* be the test scores on anchor test form A. Let $F_{XT}(x)$ and $F_{YT}(y)$ be the CDFs of test forms X and Y, then the equipercentile equating is defined as:

$$\varphi_Y(x) = F_{YT}^{-1}\big(F_{XT}(x)\big) \qquad (1)$$

where $F_{XT}(x) = \int F_P\,(x|a)dF_{AT}(a)$ and $F_{YT}(y) = \int F_Q\,(y|a)dF_{AT}(a)$. Frequency estimation works best if the two populations are similar, as it tends to give biased results when there are large group differences (Powers & Kolen, 2014; Wang et al. 2008). When there are large population differences, other methods are preferable (Kolen & Brennan, 2014, p. 146). Note that the standard errors of equating are lower for frequency estimation than for chained equipercentile equating in the NEAT design (Wang et al., 2008). Frequency estimation in the NEAT design is labelled as FE in the later result section.

Chained equipercentile estimation

In chained equipercentile equating (Dorans, 1990; Livingston et al, 1990), the CDFs F_P from test form X in population P are connected to the CDFs F_Q in population Q of test form Y through the CDFs H_P and H_Q of the anchor test forms in populations P and Q, respectively. Chained equipercentile equating can thus be defined as follows:

$$\varphi_Y(x) = F_Q^{-1}(H_Q(H_P^{-1}(F_P(x)))) \qquad (2)$$

Chained equipercentile equating is less computationally intensive than frequency estimation. If two groups of test takers differ substantially, chained equipercentile equating tends to give more accurate results in terms of a smaller bias than frequency estimation (Wang et al., 2008). If the two populations are similar and if the scores on the anchor test form and the test forms used are perfectly correlated, the results obtained from chained equipercentile equating and equipercentile frequency estimation methods are similar (von Davier et al., 2004b). This method is referred to as CE in the later result sections.

Post-stratification kernel equating and chained kernel equating

Kernel equating (von Davier et al., 2004a) comprises five steps: (1) *Pre-smoothing* the score distributions. (2) *Estimation of score probabilities* from the chosen model in step 1 (3) *Continuization* of the discrete score distributions in step 2. This involves the use of a continuous random variable, which characterizes the chosen kernel (for example Gaussian, logistic, or uniform) to be used, and a bandwidth parameter, which controls the degree of smoothness in the continuization. In this paper, the Gaussian kernel was used. The continuized CDFs of X and Y are defined respectively as $F_{h_X}(x)$ and $F_{h_Y}(y)$, where h_X and h_Y are the bandwidths. The bandwidths can be selected with different methods (see e.g. von Davier, et al., 2004a; Häggström & Wiberg, 2014) and here we used the penalty function described in von Davier et al. (2004). (4) *Equating*. The actual equating is carried out. Post-stratification kernel equating (KP) and chained kernel equating (KC) are defined respectively as

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

and

$$\varphi_Y(x) = F_{h_Y}^{-1}(H_{h_Y}(H_{h_X}^{-1}(F_{h_X}(x)))) \qquad (4)$$

where H_{h_Y} and H_{h_X} are the continuized CDFs for the anchor test forms taken by the group who took test form X or test form Y, respectively. (5) *Evaluating the equating transformation* In the final step, the equating transformation is evaluated using different accuracy measures such as the standard errors of equating, the percent relative error, standard errors and mean squared errors (Wiberg & González, 2016). An advantage with KP or KC instead of the comparable traditional equating methods frequency estimation or chained equipercentile equating, is that they give easier access to some of the accuracy measures.

A College Admissions Test

In the later empirical study, we used real test data from several test forms from the Swedish Scholastic Aptitude Test (SweSAT), which is a college admissions test given twice a year. As the spring administration is given the first part of the year, it is labelled A, and the fall administration, which is given the second part of the year, is labelled B. The SweSAT is a multiple-choice paper and pencil test with 160 binary-scored items divided into a quantitative or a verbal section with 80 items each, which are equated separately. In this paper, we only used the quantitative section, which contains subsections covering data sufficiency, mathematics, quantitative comparisons, diagrams, tables, and maps. At each SweSAT administration, a smaller sample of test takers are also given the same 40-item external quantitative anchor test while the rest of the test takers get 40 pretesting items. The quantitative part contains two sections with 40 items each. These two sections are built from the same test specifications. Also, the 40-item anchor test is built on the same test specifications as the two real test sections. The obtained test score from the SweSAT can be used by the test takers during five years, and the test takers can repeat the test as many times as they want, as only the highest score is used when they apply to college. Note, the test scores are assumed to be independent between test administrations. In the empirical study it is, however, unlikely that there are any repeaters in the sample who took the anchor test form, because the anchor test form is administered in different cities at each administration.

because the anchor test form is administered in different cities at each administration.

In Table 1, descriptive statistics for four administrations of the quantitative section of the SweSAT and the quantitative external anchor test form are displayed. From Table 1 it can be seen that the amount of test takers who took the test forms can vary a lot (from 40,431 to 76,094), which is largely due to the current situation of society with regards to unemployment rates and the general job market. The largest cohort of the SweSAT test takers were high school seniors. In the past, one used to assume that test takers who took different SweSAT administrations were equivalent and thus in the past one only used the EG design. From Table 1, it is evident that the average anchor test scores varied over the examined administrations. One possible explanation for the higher average anchor test scores for 11B could be that the unemployment rate was quite low in Sweden in 2011 and thus fewer test takers needed to take the SweSAT, compared with 2014 when the unemployment rate was a bit higher and thus more people needed a SweSAT score to apply to university programs. In other words, our examined groups were in reality not completely equivalent if one compared their anchor test score results. The standard deviation and the skewness of the anchor test scores were, however, similar over the four administrations.

Empirical Study

We examined the following four administrations of the SweSAT; 15B, 14B, 14A and 11B. Three fall administrations (labelled B as they were given at the second half of the year) were chosen as test takers who participate in the fall administrations tend to be more similar compared with test takers who participate in the spring administrations (labelled A as they were given at the first half of the year). We also included one spring administration, in order to examine how that would affect the equated values and standard errors, i.e. violation of rule 2 of linkage plans. Note, all the four examined test forms were valid at the same time for students to apply with, in order to go to college, as SweSAT results are valid for five years. We used the NEAT design as that is how SweSAT equates the test in

Table 1. Means, standard deviations, and number of test takers of the total scores and the external anchor scores of four administrations of the SweSAT quantitative section.

		Total scores			External anchor scores			
Adm	Season	M	SD	*N*	M	SD	Skewness	*N*
11 B	Fall	37.91	13.43	40,431	18.40	6.55	.36	5,263
14 A	Spring	38.28	12.72	76,094	16.37	6.32	.57	2,016
14 B	Fall	42.52	13.31	58,840	16.64	6.62	.48	2,783
15 B	Fall	42.90	12.54	60,008	17.37	6.11	.25	1,052

Adm = Administration, *N* = Number of test takers, M = Mean, SD = Standard deviation.

practice, but for comparison we also used the EG design as that was used in the past to equate the SweSAT. Note that the external anchor test is the same in all these four administrations.

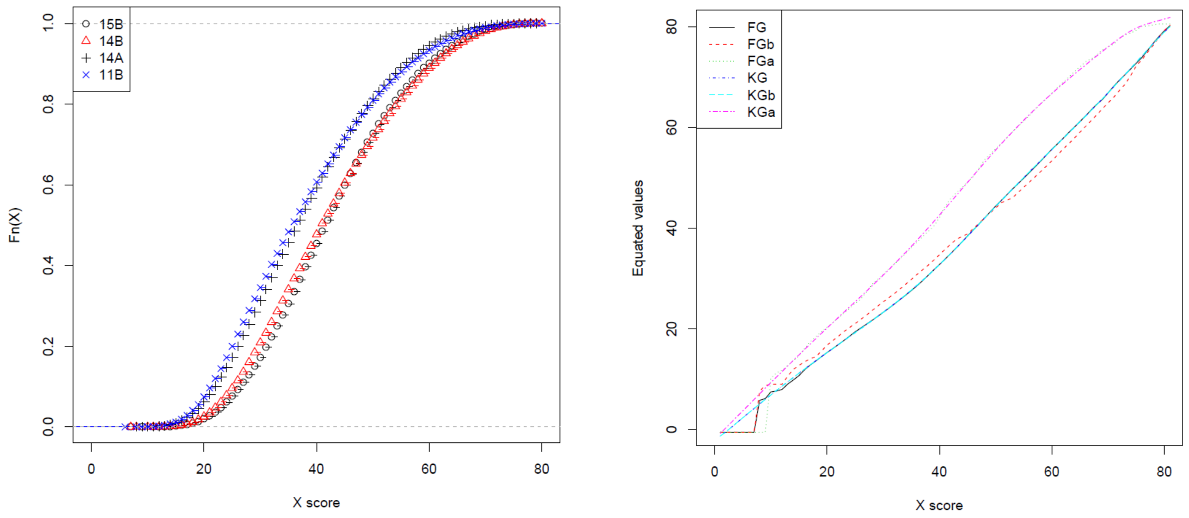
We examined the following equating methods: frequency estimation with EG design (FG), Chained equating (CE), kernel equating with EG design (KG), frequency estimation with NEAT design (FE), kernel equating with post-stratification (KP), and chained kernel equating (KC). For the kernel equating methods, we used a Gaussian kernel, and the penalty method to choose the bandwidths and the weight in the KP method was set to 0.5. The four administrations used were taken from a time series of nine test administrations, thus a large number of possible linkage plans could have been used. We chose, however, to use three different linkage plans in this paper to illustrate what happens if we are using none, one, or two intermediate test forms when equating test scores. In the first linkage plan, we equated from test form 15B directly to test form 11B. We call this linkage plan *direct* and use no extra label when it is used in the later figures. In the second linkage plan, we first equated test form 15B to 14B, and then we equated test form 14B to 14A, and finally we equated test form 14A to 11B. This linkage plan is labelled linkage plan *a* as it contained one spring administration and this letter is attached to the method names in the latter figure when this method is used. Linkage plan *a* violates the first linkage plan rule, as a longer chain is used than necessary and it also violates the second linkage plan rule as it equates test forms given at different time points. In the third linkage plan, we first equated test form 15B to test form 14B, and then we equated test form 14B to test form 11B. This linkage plan is labelled *b* as it only contains fall administrations and this letter is attached to the method names in the latter figure when this method is used.

We used plots to compare the CDFs in order to determine if they were similar over test administrations. When comparing the equated values we used the difference that matters (DTM) criterion (Dorans & Feigenbaum, 1994), which means that score differences larger than |0.5| are of concern. As there is no true equating transformation or true linkage plan, there is no absolute DTM for each method. Instead we used tables of equated values to examine if there were DTM at any of the test scores when different equating methods or different linkage plans were used. We conducted all analyses both for the empirical study and the later simulation study in R and used the R package *equate* (Albano, 2016) to perform frequency estimation and chained equating, and the R package *kequate* (Andersson et al., 2013) to perform the different kernel equating methods.

Results from the empirical study

The four test forms CDFs are given in the left part of Figure 1. Note, the CDF of test form 14A and test form 11B appear to be quite similar and the CDFs of test form 15B and test form 14B appear to be similar. From the right part of Figure 1 and the upper left part of Table 2, where the equated values are given for the different equating methods in the EG design, it is obvious that the linkage plan *direct* equating and linkage plan *b* gave similar results regardless of the equating

Figure 1. CDFs of administrations 11B, 14A, 14B, and 15B to the left and to the right the equating transformations of the two equating methods with the EG design (FG, KG) when using the three different linkage plans (*direct*, *a* and *b*). The linkage plan *direct* (i.e. equating directly from test form 15B to 11B) has no extra label, linkage plan *b* has the label "b" attached to the method name (i.e. equating from test form 15B to 11B via 14B), and linkage plan *a* has the label "a" to the method name (i.e. equating from test form 15B via 14B and 14A to 11B).



method used, while using linkage plan *a* gave different equating values. Note, both methods using linkage plan *a* gave similar equated values, although different equated values compared with the equated values from the other linkage plans. The jump on the curves in the lowest score range was probably due to the fact that there were no observed test scores in that score range.

From the upper left part of Table 2, we see for the EG design that frequency estimation (FG) gave different results if the *direct* linkage plan was used as compared with using linkage plan *b* (i.e. via test form 14B (FGb)) or linkage plan *a* (i.e. via test form 14B and 14A (FGa)). The differences were, however, larger for lower score values. When using linkage plan *b,* the kernel equating versions and the frequency estimation gave similar results, especially in the upper score range, although there were several scores with DTM.

In Figure 2 and the upper right part and lower part of Table 2, the equated values are shown when using the three different linkage plans for the NEAT design. The linkage plan *direct*, i.e. when we equate directly between 15B and 11B, is shown in both parts of Figure 2 and equating with linkage plan *a* is shown in the left part and equating with linkage plan *b* is shown in the right part of Figure 2. The kernel methods gave similar equated values when either linkage plan *direct* or linkage plan *b* was used for the higher score values – especially for score values of 60 and above. All methods, except FE for linkage plan *a* and *b* and CE for linkage plan *b*, yielded low equated values for the lower score range. For all methods, linkage plan *a* gave very different equated values compared with using either linkage plan *direct* or linkage plan *b*. Regardless of linkage plan and method, several of the equated values exhibited DTM in the mid-score range. Linkage plan *a*, however, displayed DTM over most parts of the score range for all methods, which differed from the other linkage plans.

The overall result for both the EG and NEAT designs, is that different equating results were obtained when different linkage plans were used, especially if the linkage plans differed in length. The equating values were closer to linkage plan *direct* when using linkage plan *b*, as compared with using linkage plan *a*. Note, the excluded values in Table 2 followed the same pattern as the displayed values in both the EG and NEAT designs and can be obtained from the corresponding author upon request.

Table 2. Raw test score values (X) and every tenth equated value for the two equating methods (FG, KG) in the EG design and the four equating methods (CE, FE, KP and KC) in the NEAT design when using three different linkage plans (*direct, a* and *b*).

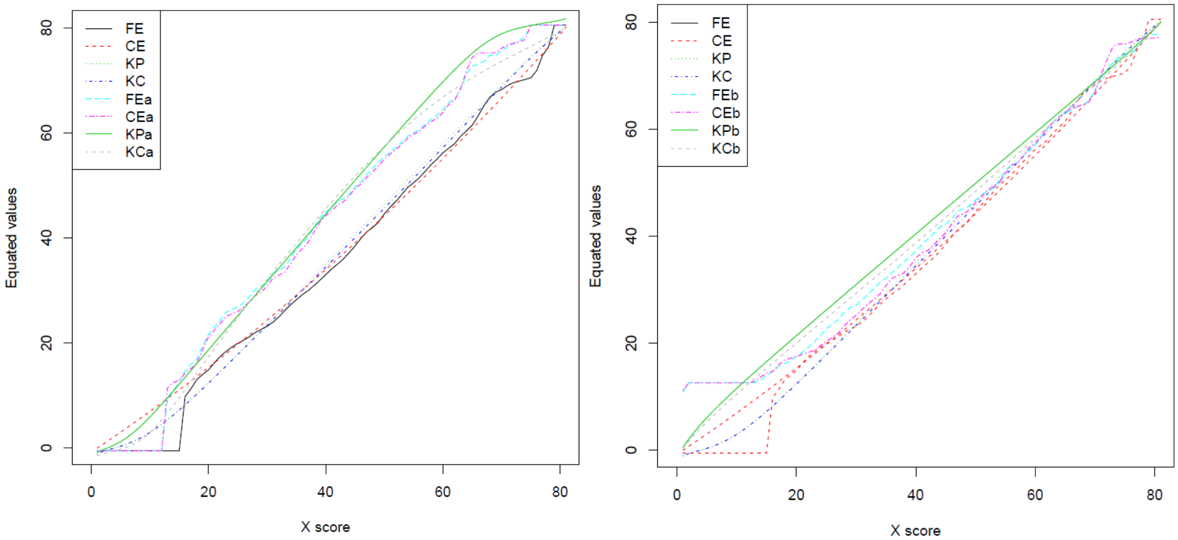
X	**FG**	**FGa**	**FGb**	**KG**	**KGa**	**KGb**	**CE**	**CEa**	**CEb**
EG design							*NEAT design*		
0	-0.5	-0.5	−0.5	−1.24	-0.71	1.02	0.00	-0.50	11.23
10	7.67	10.26	9.03	6.90	10.46	10.35	7.79	-0.50	12.58
20	16.09	20.34	17.67	15.31	21.24	18.16	16.22	22.45	17.92
30	23.31	30.86	26.21	23.31	30.70	27.34	25.29	32.37	26.32
40	33.87	42.14	35.08	32.76	43.90	36.46	34.98	45.19	36.87
50	45.50	55.80	45.03	44.31	56.74	44.68	45.29	55.98	47.37
60	56.74	66.78	54.52	55.61	67.72	53.69	56.22	65.39	58.69
70	66.86	75.96	65.98	66.87	76.83	64.78	67.79	76.75	70.14
80	80.16	80.50	80.24	79.97	81.81	77.89	80.00	80.50	77.11
NEAT design									
X	**FE**	**FEa**	**Feb**	**KP**	**KPa**	**KPb**	**KC**	**KCa**	**KCb**
0	-0.50	11.86	10.87	−1.02	-.81	0.56	−1.02	-1.43	0.39
10	-0.50	12.87	12.67	3.78	7.12	12.58	3.72	4.10	11.44
20	15.15	17.93	18.29	13.63	20.01	22.38	13.45	18.87	20.96
30	24.09	27.39	28.10	24.81	32.98	31.91	24.46	32.14	30.28
40	33.83	36.96	38.39	36.20	45.83	41.38	35.58	46.83	39.69
50	45.00	44.82	47.76	47.65	58.58	50.82	46.81	57.74	49.38
60	56.95	56.52	58.63	59.07	70.87	60.28	58.34	67.59	59.28
70	69.52	65.32	67.70	70.31	79.27	69.82	69.86	74.18	69.23
80	80.50	73.07	78.07	80.18	81.70	79.92	80.13	80.24	79.80

FG = Frequency estimation with EG design, KG = Kernel equating with EG design. CE = Chained equating, FE = Frequency estimation with NEAT design, KP = Poststratification kernel equating, KC = Chained kernel equating, b = linkage plan *b* is used, a = linkage plan *a* is used. No extra label = linkage plan direct is used.

Simulation study

To further examine the use of different linkage plans, we conducted a simulation study for the simplified case when we have three test forms; X, Y, and Z. One possibility is to equate X to Z (direct equating), another possibility is to equate X to Z via test form Y, labelled "v" to represent via in the tables and figures. To be able to connect our simulated results with the empirical study with college admissions test data we decided to mirror the real college admission test data. Thus, we sampled cases from the real college admissions test data from three test administrations. We sampled 5,000 test results from the real college admissions test data for each test form for each condition and used 500 replications. As the number of items in the real test is 80 items and 40 external anchor items, we used a NEAT design with 80

Figure 2. The equating transformation for the four equating methods with the NEAT design (FE, CE, KP and KC) when using the three different linkage plans (*direct*, *a* and *b*). The linkage plan *direct* is used in both plots and linkage plan *a* is used in the left plot and linkage plan *b* is used in the right plot.



items and 40 external anchor items. For comparison, we also examined an EG design with 80 items.

To create the condition of an easier test form, we added two score points to each test taker's sum score on that test form. To create the condition of more able test takers taking a specific test form, we added two score points to each test taker's anchor sum score. The choice of using two sum score points instead of one is due to the fact that it is typically how much higher a sum score is needed to get a higher scale score, i.e. the score that is used when applying for a university program. On the rare occasion that a test taker's sum score became higher than the maximum test score of the test, the test score was truncated to the maximum test score.

In the EG design, we used the FG and KG equating methods. With the NEAT design, we used the FE, CE, KP, and KC equating methods. For the kernel methods, we used a Gaussian kernel and the penalty bandwidth selection method and we set the weight to 0.5 in the KP method. We used a quadratic (second-order) polynomial model with one interaction term as a pre-smoothing model for the NEAT design and a simple quadratic model for the EG design. These models were chosen because we followed the principle of parsimony and these models displayed a good fit. We are aware that when conducting equating for large-scale assessments it is better to try different models and use the best fitting model. These models were, however, chosen here to limit the examined conditions. We compared the two linkage plans in the following eight conditions;

1. NEAT: Baseline case with 80 items per test form and 40 external anchor items.
2. NEAT: Easier test form X (Average of two score points higher on test form X.)
3. NEAT: Easier test form Y (Average of two score points higher on test form Y.)
4. NEAT: More able test takers taking test form X (Average of two score points higher on the external anchor test form at time point 1.)
5. NEAT: More able test takers taking test form Y (Average of two score points higher on the external anchor test form at time point 2.)
6. EG: Baseline case with 80 items per test form.
7. EG: Easier test form X (Average of 2 score points higher on test form X.)
8. EG: Easier test form Y (Average of 2 score points higher on test form Y.)

Evaluation tools

To evaluate our simulation study, we focused on the standard error (SE). To connect our simulation study with the empirical study with real college admissions test data, we also compared the equated values from the different methods and linkage plans in order to conclude if any of their score values had DTM with respect to the other used methods and linkage plans. We used replicated data generated from the real college admissions test data and compared the estimated equated values with the estimated true equated value at each test score. Let x_i denote a specific test score, where $i = 0,\ldots,n$ and the equated value $\varphi_Y(x_i)$ over *R* replications where each replicate is denoted by *r*. The SE is in general defined as $SE\left(\hat{\varphi}_Y(x_i)\right) = \sqrt{Var(\hat{\varphi}_Y(x_i))}$, and we estimated it from our data using

$$\mathrm{SE}(\varphi_Y(x_i)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\varphi}_Y^{(r)}(x_i) - \bar{\hat{\varphi}}_Y(x_i))^2} \qquad (5)$$

where $\hat{\varphi}_Y^{(r)}(x_i)$ is the estimated equated score for the r^{th} replication and the estimated true score values were calculated as $\bar{\hat{\varphi}}_Y(x_i) = \frac{1}{R}\sum_{r=1}^{R}\hat{\varphi}_Y^{(r)}(x_i)$.

Results from the simulation study

NEAT design

In Table 3, every tenth equated value of the baseline case is visible for the four equating methods using the two different linking plans for the NEAT design. Note, the omitted score values follow the same pattern as the displayed values. The simulated results in Table 3 were in line with the empirical results, i.e. using different linking plans have different impact on the equated values and that there are DTMs at many of the score values. It is also clear that different equating methods yields different equated values. The only method with similar equated scores regardless of used linkage plan was KP in the upper score range. A noticeable difference is that as we used pre-smoothing with the kernel methods we have equated scores also in the lowest score range were there were no observed score values.

The different equating methods performance on the equating transformation using the other four examined conditions within the NEAT design can be seen in Figure 3 and their SE can be seen in Figure 4. In line with the results from Table 3, from Figure 3 it is clear that using different linking plans and using different equating methods have quite a large impact on

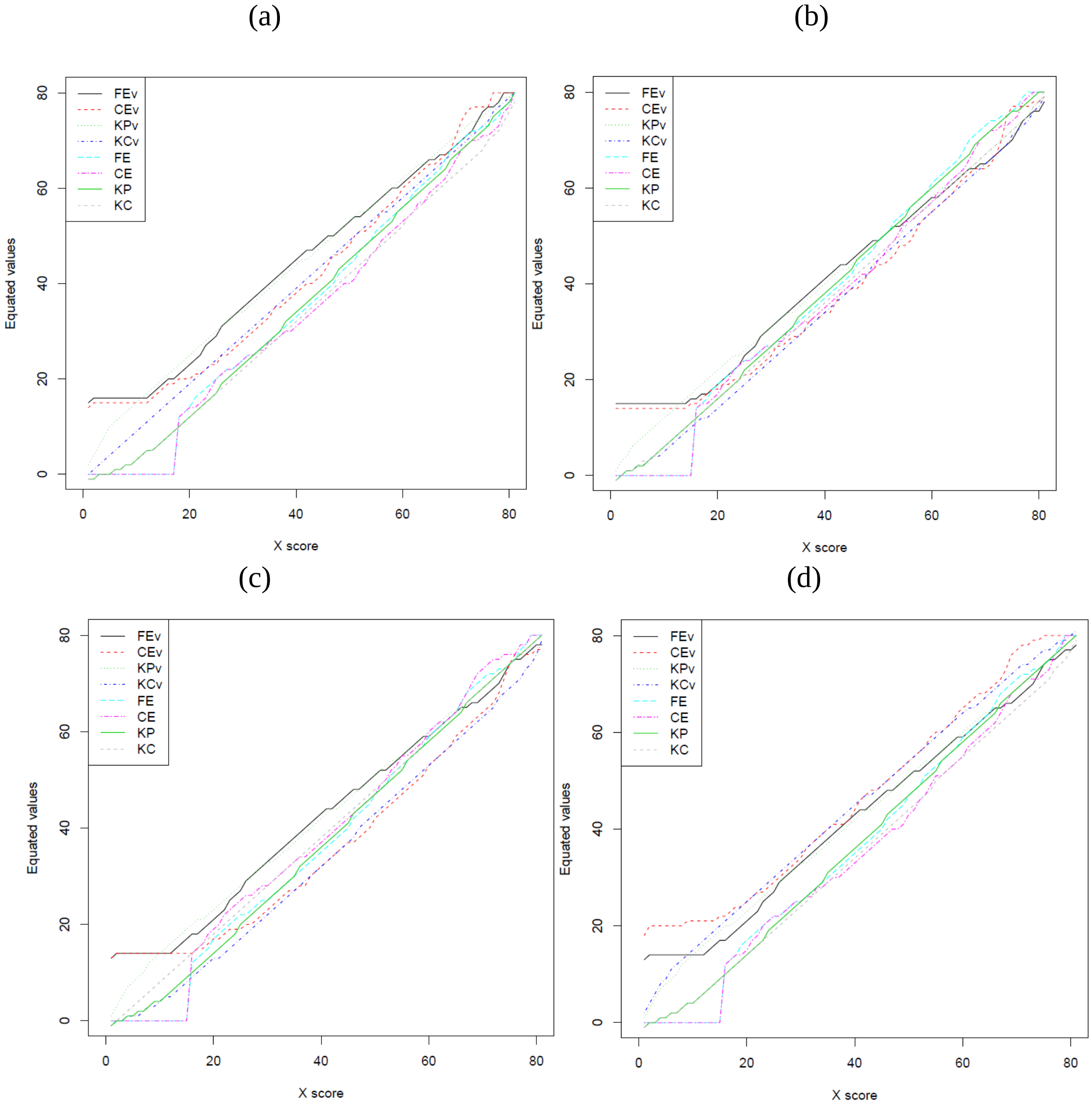
Table 3. Raw test score values and rounded equated test scores for every tenth equated value using the NEAT design for the baseline case in the simulation study. The method has the label ”v” attached when test form X was equated to test form Z via test form Y, otherwise the equating is direct from test form X to test form Z.

Score	FE	FEv	CE	CEv	KP	KPv	KC	KCv
0	0	0	0	0	0	0	0	0
10	0	0	0	0	5	0	5	8
20	18	22	16	19	15	19	15	17
30	26	34	25	27	26	27	25	27
40	37	44	36	37	37	37	35	37
50	48	51	44	47	48	47	45	48
60	60	60	58	57	61	57	58	56
70	68	72	72	70	70	70	68	66
80	78	80	77	80	80	80	80	79

the equated values. The largest impact on the equated values were observed when test form X was easier or if the test takers were more able at the first or the second time point, as can be seen in Figures 3a, 3c, and 3d. There were also DTM between the same methods with different linkage plans at some of the score values in all examined conditions. The largest observed differences were when test form X was easier or when we had more able test takers at time point 2.

From Figure 3 it is clear that CE yielded the largest SEs over most of the score range and KP yielded the lowest SEs regardless of which linking plan was used. FE and KC yielded stable SEs regardless of which linking plan was used and the size of the SEs were quite low over the score range, although slightly higher than the SEs for KP. All kernel methods had smooth SEs as expected as we included a pre-smoothing step. Slightly higher SE values were seen when an easier test form Y was used, but overall the size of the SEs was similar in the examined conditions.

Figure 3. The equating transformation for the NEAT design (a) case 2: Test form X easier, (b) case 3: Test form Y easier, (c) case 4: More able test takers at time point 1 (d) case 5: More able test takers at time point 2. The label "v" means that test form X was equated to test form Z via test form Y, otherwise the equating is direct from test form X to test form Z.



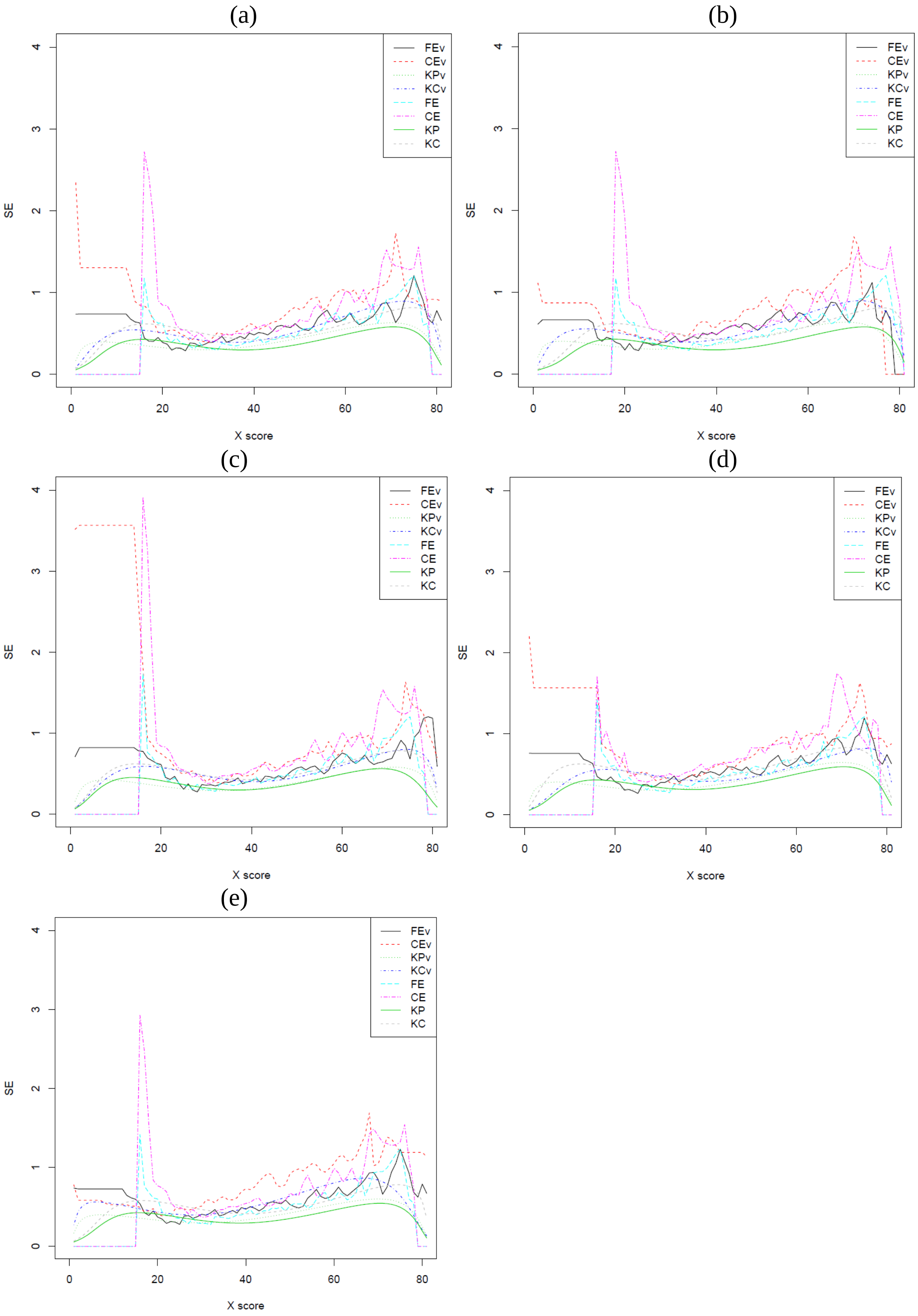
(a)

(b)

(c)

(d)

Figure 4. SEs for the NEAT design using (a) the baseline case 1, (b) case 2: test form X easier, (c) case 3: test form Y easier, (d) case 4: More able test takers at time point 1, and (e) case 5: More able test takers at time point 2.



The lowest SEs were observed for all methods in Figure 4c, i.e. when test form Y was easier. Note, the size of SE for the other four conditions was quite similar. The SE was highest on the upper score range and quite high in the lowest score range in all examined conditions as only a few test scores were observed in the upper score range and no test scores were observed in the lowest score range. Summing up, CE yielded the highest SE regardless of the linking plan used, FE and KC gave quite similar results and KP yielded the lowest SE in all examined conditions regardless of which linking plan was used.

EG design

The equating transformation for the three examined conditions can be seen in the left part of Figure 5. Similarly as in the empirical study, several test scores displayed DTM. The largest difference between the equating transformation is when test form X was easier. In general, KG has more similar equating transformations and different linking plans are used as compared with FG, which had more DTM in the test scores.

The right part of Figure 5 displays the SE for the three conditions in the EG design. A similar pattern was seen as for the NEAT design, i.e. that the kernel method has lower SEs over the whole score scale as compared with the non-kernel FE method. The SEs were reasonable low over the score scale for both methods and for both linking plans, although the SEs were slightly larger for high and low test scores, especially for the FE method. The SE was only slightly higher in the baseline case compared with the other two conditions. For the lowest score values where there are no observed test scores, the SE was high for FGv and somewhat high for FG.

Discussion and some concluding remarks

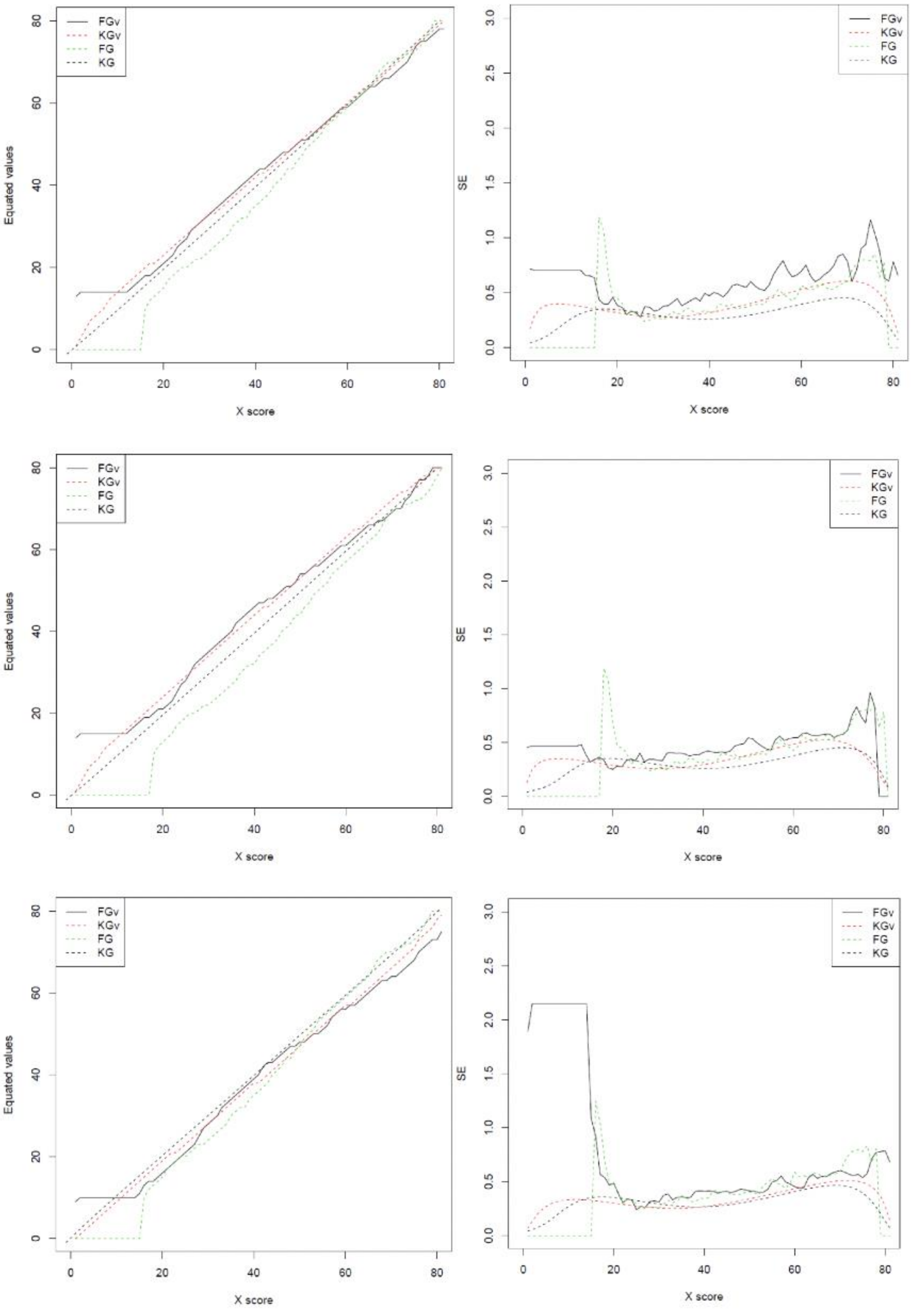
It is of utmost importance that high-stakes standardized tests be fair for different test takers who have taken different test forms at different time points. An important aspect is thus to assure that we are using equating methods and linkage plans which are stable over time and thus give fair test scores. In this study, we first conducted an empirical study using the SweSAT with three different linkage plans. As the SweSAT is currently equated with an external anchor test that we had access to, we used a NEAT design. As the SweSAT was equated with an EG design in the past, this design was also included even though the EG assumption of equivalent groups has been shown to be violated over administrations in the SweSAT (Lyrén & Hambleton, 2011).

Empirical study

One conclusion from the empirical study, was that in general, regardless of method and design, the equated values were more similar when linkage plan *direct* and linkage plan *b* were used, than when equating was done in a longer chain and at different time points as in linkage plan *a*, i.e. when violating rules 1 and 2 of linkage plans as described in Kolen and Brennan (2014). There were considerable DTM for several of the test scores, especially in the NEAT design, although it was also shown for many test scores in the EG design. A possible reason for the observed differences when different linkage plans are used is that the CDF of test form 14A was more different than the CDF of test form 14B and test form 15B. This result gives empirical evidence that one should not violate linkage plan rule 1 or 2 as dictated by Kolen and Brennan (2014). Instead, we should strive to have short equating chains and compare similar test takers, i.e. test takers who take the test at similar time points and not as in linkage plan *a* which used an unnecessarily long chain with both a fall administration (14B) and a spring administration (test form 14A) to equate test form 15B to test form 11B.

In the EG design, although linkage plan *a* yielded very different equated values compared with the other linkage plans, the two examined equating methods had similar equated values in linkage plan *a*. One reason might be that the groups were quite similar. In the NEAT design, the kernel methods yielded similar equated values when either linkage plan *direct* or linkage plan *b* was used for the higher score values – especially for score values of 60 and above. This is important, as having stable scores in the upper score scale is crucial in a college admissions test and it should not matter which equating methods are used. All methods, except FE for linkage plans *a* and *b*, yielded low equated values for the lower score range. This is probably due to the fact that we used pre-smoothing with log-linear models for the kernel equating methods and those models managed to model the lower

Figure 5. The equating transformation to the left and SEs to the right for the EG design for two equating methods (FG and KG) and two different linkage plans. First row: case 6: the baseline. Second row: case 7: easier test form X. Third row: case 8: easier test form Y. The label ”v” indicates that test form X was equated to test form Z via test form Y.



score range even though there were no observations in that score range.

Simulation study

In order to examine different linkage plans in different conditions, we included a simulation study based on the real college admissions test data with different conditions so the test forms at either time points 1 or time point 2 were easier, and that the test groups were more or less able. Similarly, to the empirical study, different linkage plans gave different results. This was true for all methods in the examined conditions, a fact that should be considered when equating test scores. One way to handle this problem is to use more than one link (Kolen & Brennan, 2014). Another way would be to use circular equating; however, Wang et al. (2008) showed that using circular equating has other problems, as it does not handle systematic errors very well.

In the simulation study, we focused on the SEs. A noticeable difference was that in the NEAT design, CE gave the highest SEs regardless of linking plan used, while FE and KC gave similar results and KP the smallest SEs over the examined conditions and regardless of which linking plan was used. The smaller SEs in the kernel methods might be due to the use of continuization and pre-smoothing in these methods. A possible reason for the larger SEs for CE might be due to increases in errors when using a chain, something which is removed when pre-smoothing is used as in KC. Thus, if the groups differ or if the test forms differ much, we suggest one use of KP or possible FE and KC.

For the EG design, the SEs was lower for the kernel method compared with the traditional method when using different linking plans. The fact that there were differences in our simulation study between the KP and CE (although FE and KC were similar) and also a difference when using the EG design is interesting, as it is not completely in line with the study of Mao et al. (2006). In that study, only trivial differences were found in a real data example when comparing the equating results of traditional equipercentile equating with those of kernel equating in the EG design and to post-stratification equating within a NEAT design. That study, however, only used real data and not simulated data as here. Our results are, however, in line with Liu and Low (2008), who also compared the use of traditional and kernel equating methods. In their study, it was evident that if groups who take two test forms are quite different from each other, then different equating methods tend to give different results – a result in line with what we found in the different explored conditions. Note, neither of those studies examined different linkage plans, nor SEs, and they did not include the KC method in their studies.

Limitations and future

A limitation with our research is that we only examined two different linkage plans, although we examined a number of different conditions and two different data collection designs. In the future one should examine if the equated values change if longer equating chains and different linkage plans are used. It would also be interesting to examine more conditions in the simulation study, including, for example, if the test forms are more or less discriminating, if different test lengths are used, and if the sample sizes are varied. If one would concentrate on the kernel methods, one could also explore the choice of different kernels and different pre-smoothing models as that may impact the equated values (see e.g. Wallin & Wiberg, 2020). In the future, one should also examine item response theory (IRT) observed-score kernel equating (Andersson & Wiberg, 2017), as that allows us to model the item with an IRT model in the pre-smoothing step.

Another interesting thing to study is the use of more than one equating transformation and to explore what happens when equating transformations are averaged by building on the research conducted by Holland and Strawderman (2011). The test in the empirical example is a college admissions test, thus the labor market has a strong influence on how many test takers and which test takers are taking the test at a given time point. When there is a recession in the economy with high unemployment, more test takers take the test since they want to start studying at colleges, in comparison to when there is low unemployment in the economy. This means that the groups who take the test can differ greatly in terms of their background over the years. In the future, one should study more closely the use of covariates and when there is a non-equivalent group with covariates (NEC) design, as Wiberg (2017) found good results with a real test data set when using kernel equating with the NEC design. This is especially a good alternative if there is no anchor test given and the test groups cannot be assumed to be equivalent (Wiberg & Bränberg, 2015) as is the case with the SweSAT.

Summing up, the different linkage plans gave somewhat different results but it depends on how similar the test taker groups are and if the test forms vary in difficulty. If one of the test forms are easier or if the ability of the test groups differs a lot, the choice of linkage plan has more impact than if the test forms or test groups are more similar. If the groups differ or if the test forms differ, it is better to use the kernel methods as they gave lower SEs at most score values. In the simulation study, the KP method in the NEAT design and the KG method in the EG design tended to give the most stable results and the fewest SEs in all the examined conditions. This result differed slightly from the empirical study results since in that study both the FG and KC methods were considered to give stable results. A reason could be that the examined groups and the test forms were quite similar in the empirical study except for 11B. In the simulation study, FG and KC had low SEs in the baseline case, when the groups were similar, and higher SEs when the groups differed. FG especially had higher SEs when the test forms differed. Thus, in practice, we recommend using KG in the EG design and KP within the NEAT design, as they exhibited low SEs in the simulation study across all the examined conditions and gave stable results in the empirical study.

References

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8),1-36.

Andersson, B., K. Bränberg, and M. Wiberg (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software,* 55(6), 1–25.

Andersson, B. & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika.* 82(1), 48-66.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement (2nd ed.)*, 508-600. Washington, DC: American Council on Education.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education*, 3(1), 3-17.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

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

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika*, *75*, 438–453.

Guo, H., Liu, J., Dorans, N. J., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift* (Research Report 11–46). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2010). *Limits on the accuracy of linking* (Research Report 10–22). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Dorans, N. J. (2011). *Sources of score scale inconsistency* (Research Report 11–10). Princeton, NJ: Educational Testing Service.

Haberman, S. J., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT I: Reasoning test score conversions* (Research Report 08–67). Princeton, NJ: Educational Testing Service.

Holland, P. W. & Strawderman, W. E. (2011). How to average equating functions, if you must. In A. A. von Davier *Statistical models for test equating, scaling, and linking.* Chapter 6, pp 109-122. New York: Springer.

Häggström, J. & Wiberg, M. (2014). Optimal bandwidth in observed-score kernel equating. *Journal of Educational Measurement, 51*(2), 201-211.

Kolen, M. & R. Brennan (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Liu, J., Curley, E., & Low, A. (2009). *A scale drift study* (Research Report 09–43). Princeton, NJ: Educational Testing Service.

Liu, J. & Low, A. C. (2008). A comparison of the kernel equating method with traditional equating methods using SAT® data. *Journal of Educational Measurement*, *45*(4), 309-323.

Livingston, S. A., & Antal, J. (2010). A case of inconsistent equatings: How the man with four watches decides what time it is. *Applied Measurement in Education*, *23*, 49–62.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3(1), 73-95.

Lyrén, P-E., & Hambleton, R.K. (2011).Consequences of violated the equating assumptions under the equivalent group design. *International Journal of Testing, 36*(5), 308-323.

Mao, X., von Davier, & Rupp, S. L. (2006). Comparisons of the kernel equating method with the traditional equating methods on Praxis data. ETS research report, RR-06-30.

Moses, T., Deng, W., & Zhang, Y. (2011). Two approaches for using multiple anchors in NEAT equating: A description and demonstration. *Applied Psychological Measurement*, *35*, 362–379.

Powers, S. & Kolen, M. J. (2014). Evaluating equating accuracy and assumptions for groups that differ in performance. *Journal of Educational Measurement, 51*, 39-56.

Puhan, G. (2009). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education*, *22*, 79–103.

Taylor, C. S., & Lee, Y. (2010). Stability of Rasch scales over time. *Applied Measurement in Education*, *23*, 87–113.

von Davier, A. A., P. Holland, and D. Thayer (2004a). *The kernel method of test equating*. New York: Springer-Verlag.

von Davier, A. A., P. Holland, and D. Thayer (2004b). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41,* 15-32.

Wallin, G. & Wiberg, M. (2019). Propensity scores in kernel equating under the non-equivalent groups with covariates design. *Journal of Educational and Behavioral Statistics. 44*(4), 390-414.

Wallin, G. & Wiberg, M. (2020). *Selecting a presmoothing model in kernel equating.* In Wiberg, M., Molenaar, D., González, J., Böckenholt, U., & Kim, S-J. (Eds.). Quantitative Psychology – 84th Annual Meeting of the psychometric society, Santiago, Chile, 2019, New York: Springer. 111-120.

Wang, T., Hanson, M. J. and Harris, D. J. (2000). The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement, 24,* 195-210.

Wang, T., Lee, W-C, Brennan, R. L., & Kolen, M. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651.

Wiberg, M. (2017). *Ensuring test quality over time by monitoring the equating transformation.* In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W-C. Wang. (Eds.) Quantitative Psychology, New York: Springer. 239-252.

Wiberg, M. & K. Bränberg (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, *39*(5), 349–361.

Wiberg, M. & González, J. (2016). Statistical assessment of estimated transformations in observed-score equating, *Journal of Educational Measurement*, *53*(1), 106-125.

Wiberg, M. & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing.* 1-22. 105-126.

.

Citation:

Wiberg, M. (2021). On the use of different linkage plans with different observed-score equipercentile equating methods. *Practical Assessment, Research & Evaluation*, 26(23). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/23/

Corresponding Author

Marie Wiberg
Umeå School of Business, Economics, and Statistics
Umeå University,
Umeå, Sweden

email: marie.wiberg [at] umu.se