CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

The biggest equating study in the world... ever

Research Report

Tom Benton
Matthew Carroll
1st December 2025

Author contact details:

Tom Benton & Matthew Carroll
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

tom.benton@cambridge.org
matthew.carroll@cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division
If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Benton, T. & Carroll, M. (2025). *The biggest equating study in the world… ever*. Cambridge University Press & Assessment.

Abstract

What is this report about?

This report presents the largest empirical evaluation of test equating methods ever conducted. Equating is the process of identifying scores on different test forms that can be treated as equivalent. Whilst various equating methods are available, there remains no consensus on which are most effective under which conditions. This study aims to provide practical guidance by comparing 69 techniques across 60 scenarios using real data from 1,000 high-stakes tests.

What did we do?

We conducted a large-scale simulation using item-level data from 1,000 real examinations administered by Cambridge OCR and Cambridge International. Each test was split into two pseudo-forms with shared anchor items, and candidates were split into two groups, one taking each pseudo-form. Equating methods were applied to estimate score equivalencies between forms, and their accuracy was assessed against a criterion equating function derived from the full dataset. The simulation varied three key conditions: sample size (50 to 2,000), anchor test length (18% to 40% test form overlap), and group ability differences (random, medium, and maximum). Accuracy was primarily evaluated using weighted mean absolute error (MAE). Two new methods were developed: odds-transform equating, and transformed linear equating. Both are described in the report and have been made available for practitioners.

What did we find?

- **Best-performing methods varied by scenario**. For small samples (≤100 candidates per test form), Rasch PCM true score equating (with JML estimation) and odds-transform methods were most effective. For large samples (≥500 candidates per test form), IRT observed score equating using GRM or GPCM models performed best with medium or large group differences, whilst Frequency Estimation Equipercentile with kernel smoothing performed best with random groups.
- **Classical methods** like kernel-smoothed equipercentile equating worked well when group differences were small and sample sizes were large. Circle-arc and odds-transform equating methods, originally designed for small samples, performed surprisingly well even with larger samples and large group differences.
- **Performance of different methods was often broadly similar** meaning that although there were clear “best” approaches, if equating has been carried out using other methods, the results should still be acceptable.
- **Several methods were consistently outperformed** by at least one other method, so could be excluded from consideration in most cases. These include identity equating, unsmoothed equipercentile equating, and certain IRT variants.
- **For IRT equating, model fit had some impact**, but the methods remained robust even when fit indices were below recommended thresholds.

What are the implications?

This study provides evidence-based recommendations for practitioners selecting equating methods. It narrows the field to a core set of reliable techniques and identifies when each should be used. Odds-transform equating emerges as a promising classical method due to its simplicity, robustness, and performance across diverse scenarios. Furthermore, given good performance in a range of scenarios, equating using the Rasch PCM (ideally with concurrent estimation) could be considered a good “default” approach. Findings highlight the importance of considering sample size and group differences when selecting methods.

Table of contents

Introduction

Within educational measurement, equating refers to the process by which we can identify pairs of scores on two test versions (or forms) that can be treated interchangeably. For example, if a score of 32 on test form X is equated to a score of 37 on test form Y, this implies that these two scores can be treated as having the same meaning. This is necessary in situations including test administrations using multiple forms, tiered tests, and maintenance of standards over time. Numerous equating methods have been suggested. This report uses real empirical data to determine which of these methods work best in different scenarios. This is an important topic for research and practice as, despite the proliferation of methods, according to Livingston (2014, p. 67) "Among people whose work includes the equating of test scores, there is still (as of 2014) no consensus on this question."

From a methodological perspective, statistical linking often refers to the same practical procedures as equating, but with a weaker form of interpretation – perhaps that two test scores are *equally hard* to achieve, but not necessarily that they represent *exactly the same* kinds of knowledge and skills. As such, the recommendations here regarding methods for equating apply equally well to methods of linking.

A detailed description of many of the methods used for equating and linking is provided by Kolen & Brennan (2004)[1]. These are broken into methods that can be used when the test forms being equated are taken by equivalent groups of candidates, and those where the groups taking different test forms differ in ability. In this latter situation, we must account for this ability difference using performance in an anchor test taken by both groups, i.e., a set of items, or even a separate test, which *both* groups take.

The former situation, encompassing designs known as the single group design, equivalent groups design and random groups design, is easily addressed by the equipercentile approach to equating. That is, we define scores on two forms as equivalent if the percentage of candidates achieving at or below each score is the same. As such, the issue of choosing an equating method reduces to the problem of how to estimate the percentiles of each score distribution. Whilst a few special techniques have been suggested for circumstances with very small samples and reasonably large differences in test forms (e.g., Livingston & Kim, 2009), the problem of how to perform equating in scenarios with equivalent groups is largely solved.

The situation where groups taking different test forms differ in ability, with an anchor test used to account for this, is generally known as the NEAT design (Non-equivalent anchor test design). The best method to address this situation is much less clear. A very large number of techniques for use with the NEAT design have been suggested. Even a cursory read of the contents pages of Kolen & Brennan (2004) suggests at least nine possible methods:

- Tucker Method
- Levine Observed Score Method
- Levine True Score Method
- Mean Equating
- Braun-Holland Linear Method
- Frequency Estimation Equipercentile Equating
- Chained Equipercentile Equating
- Equating True Scores (from IRT[2]).
- Equating Observed Scores (from IRT).

[1] As well as more recent editions of the same book.

[2] IRT stands for item response theory.

However, for several of the above techniques, subvariants are available, meaning that the total number of approaches is far higher. For example, each method of equipercentile equating can be applied using pre-smoothing, post-smoothing, or no smoothing. Any method based upon IRT can be built upon any of the various models (e.g., the Rasch model, the graded response model, the 3-parameter model, etc.; see, for example, de Ayala, 2013). These considerations alone mean that we are faced with dozens of methods, with little guidance on which is best[3].

Beyond Kolen & Brennan's book, authors have proposed various new and alternative approaches to equating. Some of these will be introduced in the methods section. However, though the research paper accompanying each new method always includes some empirical work showing the effectiveness of the method, these studies can be limited in at least two ways:

1. The performance of any newly proposed technique is only compared to a fraction of those proposed previously. To take one, by no means unusual, example, when Wang & Brennan (2009) proposed their modified frequency estimation approach to equating, they only compared it to (unmodified) frequency estimation and to chained equipercentile equating. Whilst restricting focus to certain methods is reasonable in the context, it makes it hard to evaluate performance against other methods more broadly.

2. Performance is often evaluated against a small number of real datasets or against simulated data. For example, Wang & Brennan (2009) evaluate their technique using simulated data[4]. The same is true of Sun & Kim's (2023) comparison of equating methods. Other studies such as von Davier *et al.*'s (2006) evaluation of kernel equating, or Livingston and Kim's (2009) paper on circle-arc equating evaluate the technique against a single real dataset, whilst others, such as von Davier & Chen's (2013) evaluation of hybrid equating, use two real datasets. This approach to evaluation may be insufficient. The scenario that can arise is to read about a technique that appears highly effective, but when applied to real data its performance is not as strong as anticipated. For this reason, we may not be able to entirely trust newly proposed statistical methods that have been trialled in very few or idealised (i.e., simulated) scenarios.

The large number of approaches to equating requires a large study. This study should be large in terms of the number of methods considered, and the number of datasets used. That is what we seek to achieve here: an evaluation of numerous equating methods, using a large number of *real* datasets. All of the datasets used in analysis will be derived from genuine examinations so that the data reflects the behaviour of test items when taken by real candidates. This means that we do not expect them to perfectly fit the assumptions of any particular underlying statistical model. In turn, this should give a realistic and practical guide to which methods work best, which methods do *not* work well, and under which circumstances methods might or might not work.

Having presented the problem we seek to address and, broadly, what we aim to do about it, we now present our attempt at undertaking the biggest equating study in the world…ever.[5]

---

[3] Table 8.5 of Kolen & Brennan's (2004) book provides some general guidance. However, it stops short of providing any clear recommendations that practitioners should use one technique rather than another.

[4] With simulations based on IRT with item parameters based upon a single real dataset. A real dataset is also analysed in the paper but only for the purposes of illustration rather than evaluation.

[5] The title is inspired by "The Best… Album in the World…Ever" music compilation albums, which the authors fondly remember from the 1990s (see https://en.wikipedia.org/wiki/The_Best..._Album_in_the_World...Ever!).

Methods

To compare equating methods, we carried out a large simulation using real test data. The overall process was to select a large number of tests, and then for each test, split item-level data into two separate pseudo-tests including several shared anchor items. The true relationship between scores on each pseudo-test was estimated using all candidates. Then, separate samples of candidates were selected for each pseudo-test, and various equating methods were used to estimate the relationship just using those candidates' scores. The resulting equating functions were then compared to the true equating function, and metrics were calculated to evaluate their performance under the various different conditions applied. Here, we describe this process in more detail.

Test data

We acquired item-level data from tests administered as part of high-stakes qualifications by either Cambridge OCR (https://www.ocr.org.uk/) or Cambridge International (https://www.cambridgeinternational.org/). For inclusion, tests were required to have at least 5,000 candidates, at least 15 items, at least one item worth 1 mark, at least one item worth more than 1 mark, no optional items, and no items showing negative discrimination. Beyond these conditions, we permitted any test to be included, regardless of qualification type, subject, year, or session. The list of eligible tests was finalised in August 2023, so included tests taken up to the summer sessions of that year. The full list constituted over 2,500 tests.

The analysis is based upon 1,000 tests randomly drawn from the full set; this number was chosen to include a wide range of test types, but to permit the simulation to complete in an acceptable timeframe. Tests came from 2009 to 2023, and from various sessions within those years. Note that there are fewer tests from 2020 and 2021 due to the cancellation of exams during the Covid pandemic, and there are more tests from June, the main examination session in England and in some other countries. In terms of subjects and qualifications, as may be expected, the greatest numbers came from those with the greatest numbers of entries (for example, Cambridge OCR GCSEs and Cambridge International IGCSEs) but other qualifications (such as A levels) were also included in analysis. In terms of subjects, around two-thirds of the analysed tests were in Mathematics and Science subjects, and about a sixth were in English with the remainder from a range of subjects including Computer Science, Geography, Accounting and various Modern Foreign Languages. The primary take-away from this, then, is that the sample contained tests from various years, months, qualifications and subjects, meaning we can be confident that conclusions drawn from the simulation are not limited to particular conditions.

Equating methods

Sixty-nine methods, or method variants, were applied in the simulation. These are listed in Table 1. The methods can be split into those using IRT equating and those based on classical techniques. For IRT equating, four different kinds of IRT models were considered with a number of variants within each one, depending upon the exact approach to model estimation and equating that was adopted. Classical equating methods break into 9 basic groups with various variants of each one. Slightly longer descriptions of each equating method can be found in Appendix 2.

Table 1. Summary of equating methods, and the specific variants, applied in the main simulation.

IRT or classical	Type	N variants	Variants
IRT	PCM	12	MML SL True, MML SL Observed, MML Haebara True, MML Haebara Observed, MML Concurrent True, MML Concurrent Observed, CML SL True, CML Haebara True, CML Concurrent True, JML SL True, JML Haebara True, JML Concurrent True
IRT	GRM	6	MML SL True, MML SL Observed, MML Haebara True, MML Haebara Observed, MML Concurrent True, MML Concurrent Observed
IRT	GPCM	6	MML SL True, MML SL Observed, MML Haebara True, MML Haebara Observed, MML Concurrent True, MML Concurrent Observed
IRT	GRM/3PL	4	MML SL True, MML SL Observed, MML Haebara True, MML Haebara Observed
Classical	Frequency Estimation Equipercentile	6	Unsmoothed, Bump smoothed, Loglin smoothed, Modified WB Unsmoothed, Kernel smoothed, Hybrid
Classical	Chained Equipercentile	4	Unsmoothed, Bump smoothed, Loglin smoothed, Kernel smoothed
Classical	Linear	6	Chained, Braun-Holland, Tucker, Levine Obs, Levine True, Post-stratification equating (PSE)
Classical	Transformed linear	2	Chained, Tucker
Classical	Mean	5	Chained, Braun-Holland, Nominal, Tucker, Levine Observed
Classical	Circle-arc	5	Chained, Braun-Holland, Nominal, Tucker, Levine Observed
Classical	Symmetric circle-arc	5	Chained, Braun-Holland, Nominal, Tucker, Levine Observed
Classical	Odds-transform	7	Chained, Chained (SD), Tucker, Tucker (SD), Levine, Levine (SD), Nominal
Classical	Identity	1	

To briefly summarise the methods in Table 1:

- Rasch partial credit model (PCM) approaches to equating were attempted in 12 different ways. This included 3 different fitting algorithms: marginal maximum likelihood (MML), joint maximum likelihood (JML) and conditional maximum likelihood (CML). Both true score methods of equating[6] and observed score methods of equating[7] were trialled. Finally, both concurrent estimation (where all data is included in a single model), and separate estimation of models in different groups followed by metric transformation to align IRT parameters (using either the Stocking-Lord or Haebara methods) were trialled.
- More complex IRT approaches were trialled using the generalised partial credit model (GPCM), the graded response model (GRM), and a version of the GRM that fitted a 3-parameter model to dichotomous items. Each of these was fitted using MML estimation. Both concurrent (using multiple-group IRT) and separate estimation approaches were trialled, as were both true and observed score equating methods.

[6] True score equating defines two scores as being equivalent if they are the expected scores on different tests at the same level of IRT ability.

[7] Observed score equating uses an IRT model to predict the overall score distribution of each test if they were taken by a common population of candidates. After this, equipercentile methods are used based upon the predicted distributions.

- Six frequency estimation (FE) and four chained approaches to equipercentile equating were trialled. Most of these differed only in terms of whether and how smoothing was applied. One, more unusual smoothing method was the kernel-based FE approach to equating. This process attempted to identify sample weights so that, for each group of candidates, the first four moments of the anchor score distribution[8] matched those of the distribution in the joint population[9]. This was achieved using the Minimum Discrimination Information Adjustment (MDIA) approach of Haberman (1984). Two additional innovative FE methods that attempt to adjust for measurement within the anchor test were also included: the modified Wang-Brennan (WB) method (Wang & Brennan, 2009) and the hybrid method of von Davier & Chen (2013).
- Six versions of linear equating were attempted. Of these, the only non-standard approach was one based upon post-stratification equating (PSE). This again used the MDIA method to match the distribution of scores on the anchor test before applying linear equating. For five of these methods, simple equivalents based upon mean equating were also trialled.
- For two of the variants of linear equating, additional versions were trialled where the scores on each form were transformed before linear equating was applied. Further details are given below.
- Two different approaches to circle-arc equating were trialled: standard (or non-symmetric) and symmetric (see Livingston & Kim, 2008). Both methods attempt to identify an equating curve that ensures (1) zero on one test maps to zero on the other, (2) the maximum possible score on one maps to the maximum possible score on the other, and (3) the mean score on one test maps to the mean score on the other[10]. For each method, 5 different approaches to identifying what the mean score would be on each test in a common population were trialled. Initial trialling found that circle-arc approaches were very successful, and so an additional method, called odds-transformed equating, was developed for inclusion in the current research. Further details are given below and also in Appendix 1.
- Finally, as a very basic comparator for all methods, identity equating, that assumes that every score on one test form maps to the same proportion of the maximum available score on the other was included.

Odds-transform equating

The development of odds-transform equating was motivated by several issues that make routine use of circle-arc equating problematic. Firstly, where the mean scores of the two forms being equated are very different, the relationship from circle-arc equating is not necessarily monotonic. That is, as the scores of form X increase, the equivalent scores on form Y might *decrease*. Secondly, circle-arc equating is not particularly easy to implement. Whilst it is implemented within the *equate* R package (Albano, 2016), coding the method from scratch into other software packages is not particularly straightforward. Obviously, this issue is not insurmountable but if an easier way to achieve the same goals was available this would be preferred. For these reasons we developed odds-transform equating. The method meets all three goals of circle-arc equating but can be implemented in a single formula and guarantees a

[8] That is, the mean, variance, skewness and kurtosis.

[9] That is, the anchor score distribution across both Group 1 and Group 2 combined.

[10] By mean score we generally mean the expected mean scores if both forms were taken by the same group of candidates.

monotonic equating relationship. In addition, it is easy to extend so that not only do the mean scores on different test forms map to one another but, at the mean scores, the gradient of the equating function is equal to the ratio of the standard deviations on the two tests[11]. Further details of the method are given in Appendix 1.

Transformed linear equating

A second idea, also inspired by circle-arc equating, was simply to apply non-linear transformations to all raw scores before applying the usual linear equating techniques. After linear equating is completed, the non-linear transformations are reversed so that we have the usual raw score to raw score equating table. The exact non-linear transformation was to replace a score of Z on any test form with a score of

$$Z^* = \operatorname{arctanh}\left(\frac{\left(\frac{Z}{max}\right) - 0.5}{0.6}\right)$$

Here, max is the maximum score available on a test form and $\operatorname{arctanh}$ is the inverse hyperbolic tangent function. This transformation was chosen as it is almost exactly linearly related to the (somewhat more complex) arcsine transformation suggested by Kolen (1988) to stabilise error variance across the score range. However, the above transformation is preferred as it has an easy analytical inverse which makes implementation much easier in the context of equating.

This has two potential advantages. Firstly, whilst the use of the transformation before equating (and its inverse afterwards to convert back to raw scores) does not *guarantee* that scores of zero and the maximum available on each form will map to one another, it will tend to make equated scores *closer* to these points than untransformed linear equating. In addition, given that linear equating involves the calculation of means and standard deviations, it may be more effective if the scores these are calculated from have the same level of measurement error.

Applying equating methods

Equating methods were applied using functions in R statistical software. Most classical equating methods were applied using the *equate* package (Albano, 2016). Most IRT models were fitted using the *mirt* package (Chalmers, 2012). This includes all GRM, GPCM, and GRM/3PL models, and the PCM model fitted via MML. PCM models using JML were fitted with the *TAM* package (Robitzsch *et al.*, 2024), while those using CML were fitted with the *dexter* package (Maris *et al.*, 2025). In all IRT methods, equating via Stocking-Lord or Haebara methods was carried out using the *unimirt* package (Benton, 2023). Hybrid equating, chained equipercentile equating with kernel smoothing, hybrid equating, odds-transform and transformed linear equating were completed using the *KernEqWPS* package (Benton, 2025). Wang-Brennan (WB) equating was completed using bespoke code developed for the purposes of this project.
For loglinear smoothing, smoothing was applied up to bivariate interactions. The degree of smoothing started at a high level (16th order polynomial for univariate interactions, and 8th order polynomial for bivariate interactions), but such high levels of smoothing could produce errors; if this happened, sequentially lower degrees were trialled until one worked (with a minimum of 4th order polynomial for univariate interactions, and 2nd order polynomial for bivariate interactions).

[11] By standard deviation, we mean the expected standard deviation on each test form if they were taken by the same group of candidates.

Simulation conditions

Three key conditions were applied in the simulation, based on aspects of the test population or test design that we considered likely to influence equating success. First, the **anchor test length** was varied. Four conditions of ⅒, ⅐, ⅕ and ¼ of total marks (i.e., marks on the original test) were used, representing relatively short to relatively long anchor tests. To make these conditions more interpretable, we hereafter refer to them in terms of *overlap*, or the proportion of each of the two test forms derived from the original test that the different anchor lengths represented (i.e., 18%, 25%, 33% and 40%). For example, if the anchor length is ⅐ of total marks, each form takes ⅐ of the total marks plus half of the remaining marks (i.e., $^3/_7$ of the total marks). Thus, in this condition, each form comprises of $^4/_7$ of the total marks of which a quarter (25%) are in common with the other form. Note that, in reality it may not be possible to select exactly the intended number of marks (for example if the total is not divisible by 7). In these cases, the anchor length is selected to be as close as possible to the length implied by the conditions[12]. For practical reasons, these four anchor length conditions were simulated separately, so specific random realisations differed from one another (e.g., when sampling items for each form at random, a different sample could have been drawn under each of the different anchor length conditions).

Next, the **ability difference between groups** was varied. To define two groups to take the different test forms, the item (or item *mark* for polytomous items) with facility closest to 50% was identified. Candidates who gained the focal mark were put into one group (labelled Group 2), and candidates who did *not* gain it were put into another (labelled Group 1). This grouping represented the "maximum difference" condition, in that it was entirely based on performance. This approach generally led to one higher-scoring and one-lower scoring group, approximately reflecting what might be expected from a tiered exam. After this, a random group assignment was created, following a binomial distribution with probability 0.5; this grouping was not based on ability, and thus produced two groups with very similar performance. This was used as the "random groups" condition, and approximately reflects what would be expected from giving two test forms to the same class of students. Finally, an intermediate condition was created by applying the random group ID to a random 50% of candidates. That is, approximately half of the candidates kept their original, performance-based grouping, and the rest received the random grouping. This was known as the "medium difference" condition, and reflects a scenario where different test forms are given to groups with overlapping, but slightly different, abilities.

The final condition applied was the number of students taking each pseudo-test, or the **sample size**. Conditions of 50, 100, 500, 1,000 and 2,000 students were used, with the required number sampled at random from the students in each group. Note that the sample sizes refer to *the number of students taking each pseudo-test*. Conditions with *mixed* sample sizes (e.g., equating a test taken by 50 students to one taken by 2,000) were not considered.

Simulation process

The full list of possible tests to use was put into a random order, with the simulation then proceeding through one test at a time. The following process was followed:

- Extract item-level data for chosen test from the institutional data store.
- Find the item/mark point that produces closest to a 50/50 split in candidates.

[12] For example, in reality, in the ⅐ condition, anchor lengths were between 13% and 16% of the total test length (with most exactly equal to either 14% or 15%).

- Split candidates into two groups based on performance in the chosen item, then discard that item from further analysis.
- Randomly select items for which the marks add up to the required anchor length (i.e., 1/10, 1/7, 1/5 or 1/4 of total marks[13]); these become the anchor items for both test forms (using *KernEqWPS* package (Benton, 2025) for random selection).
- Randomly split the remaining items into two forms, aiming to get as close as possible to 50% of the remaining marks in each (using a function within the *KernEqWPS* package (Benton, 2025) for random selection)[14].
- Combine unique items and anchor items into form X and form Y (each including an internal anchor).
- Carry out criterion equating, using unsmoothed equipercentile equating, between the two forms, using data from all students. This is justified as each of the original datasets had a large number of candidates (at least 5,000) and so the unsmoothed approach should be very accurate. Furthermore, the equipercentile definition of equating is the one that would be most defensible to users of test results (e.g., teachers) if two test forms had been taken by equivalent groups of candidates.
- Split the candidates into two groups according to the focal condition (i.e., based on performance in the splitting item, based on random assignment, or a 50/50 mixture of the two grouping methods).
- From each group, randomly select the required number of candidates for that iteration (i.e., 50, 100, 500, 1,000 or 2,000).
- From the full item-level dataset, extract data for items in Form X for Group 1, and items in Form Y for Group 2.
- Apply all equating methods to the two datasets, and store the full equating function.
- Loop through all conditions for a given dataset before moving to the next; proceed until the required number of tests has been completed.

The four different anchor length conditions were run in separate instances. This means that some caution should be applied when comparing between anchor length conditions, as it was not the case that the longer length anchors included all or even any of the items in the shorter anchor length condition. However, *within* each anchor length condition, comparisons between sample sizes and group differences can be consistently made.

The simulation process meant that there were 60 combinations of conditions, hereafter referred to as “scenarios”. That is, there were 4 anchor lengths, 5 sample sizes, and 3 group differences, which, multiplied together, gives 60 scenarios.

Description of simulated scenarios

The process described above produced 1,000 realisations of two test forms under each of four different anchor length conditions. There was therefore extensive variety in numbers of items, numbers of marks, and difficulty. Table 2 provides descriptive statistics for these quantities, to give an understanding of simulated test forms.

[13] Excluding the item used to split the data into groups.

[14] Note that forms X and Y are not guaranteed to have the same number of marks (e.g., if the original total number of marks is odd). However, the maximum available scores on the two forms never differed by more than 1.

From the table, we see that simulated forms were around 17 to 20 items long on average, with Form X being one item longer than Form Y on average. There was substantial variation though, with the smallest forms being in the region 6 to 9 items long, and the longest being 36 to 42 items long. Marks followed a broadly similar pattern, with means ranging around 36 to 42, but with Form Y showing a slightly higher mean. The smallest numbers of marks in a form ranged from 15 to 17, whilst the largest number ranged from 75 to 86.

Form difficulty, estimated across *all* candidates (not just those sampled within the simulation) was well balanced, with an average difference in mean scores close to 0 in all cases. Again, however, variation was evident, with minimum and maximum difficulty differences of 7 to 9 marks in both directions.

Table 2. Descriptive statistics for the different test forms simulated. Numbers of items and marks relate to the full form, i.e., form-specific items *plus* anchor items. Mean scores are estimated from the *full* population of candidates, i.e., all candidates who took the real test, *not* just those sampled as part of the simulation.

Form overlap		Form	Minimum	Mean	Median	Maximum
18%	N items	X	8	18.2	17	38
		Y	6	17.0	16	36
	N marks	X	15	35.8	32	75
		Y	15	36.3	33	76
	Mean score	X	5.6	19.9	18.3	56.5
		Y	5.9	20.1	18.7	54.2
		Difference (Y-X)	-8.3	0.2	0.2	7.8
25%	N items	X	8	19.0	18	39
		Y	6	17.7	17	38
	N marks	X	15	37.3	34	78
		Y	16	37.7	34	79
	Mean score	X	5.6	20.8	19.0	56.5
		Y	5.3	20.8	19.2	56.6
		Difference (Y-X)	-7.4	0.0	0.0	8.3
33%	N items	X	9	19.7	19	41
		Y	7	18.6	18	40
	N marks	X	16	39.1	35	82
		Y	17	39.6	36	83
	Mean score	X	6.6	21.7	19.8	58.5
		Y	6.4	21.9	20.1	61.7
		Difference (Y-X)	-8.1	0.2	0.2	9.3
40%	N items	X	9	20.5	20	42
		Y	6	19.3	18.5	41
	N marks	X	17	40.8	37	86
		Y	17	41.2	37	86
	Mean score	X	6.1	22.7	21.1	62.8
		Y	7.5	22.9	21.2	62.1
		Difference (Y-X)	-8.6	0.1	0.1	9.4

Similarly, we can also examine differences between the groups of candidates used in the simulation. Three group conditions were used: “maximum difference” (purely deterministic, based on performance on the splitting item); “medium difference” (half deterministic and half random); and “random” (all random group assignment). Differences between these different groups, are summarised in Table 3. Values in the table refer to the mean marks obtained by *all* candidates in each group (i.e., not just those sampled on specific iterations), on the total score

across all items in the original dataset[15] (i.e., those in form X and form Y combined) and the standardised difference between them (in SD units).

Table 3. Summary statistics for test scores and standardised differences from all candidates in the different group conditions.

Group difference		Minimum	Mean	Median	Maximum
Maximum	Group 1 (marks)	8.7	31.0	28.5	86.9
	Group 2 (marks)	13.6	41.3	37.7	110.7
	Difference (SDs)	0.1	0.8	0.8	1.4
Medium	Group 1 (marks)	10.1	33.6	30.8	92.5
	Group 2 (marks)	12.8	38.8	35.6	104.8
	Difference (SDs)	0.0	0.4	0.4	0.7
Random	Group 1 (marks)	11.7	36.2	33.4	98.5
	Group 2 (marks)	11.6	36.2	33.4	98.5
	Difference (SDs)	-0.1	0.0	0.0	0.1

From this table, we see that the maximum group difference condition related to, on average, a difference in performance of a little under 1 standard deviation, with Group 2 (those who gained the mark on the splitting item) showing higher attainment overall. In this condition, mean performance was around 10 marks higher for Group 2. In the medium difference condition, the difference between mean marks was reduced to around 5 on average, and the average standardised difference was a little under 0.5 standard deviations. As may be expected, in the random grouping condition, mean marks were identical on average, and the standardised difference was 0 standard deviations.

Evaluating equating accuracy

Equating functions for every condition and every test were stored during the simulation. Accuracy was then assessed on completion of the simulation using three metrics: the mean absolute error (MAE), bias, and root mean square error (RMSE). All of these were calculated as weighted forms, with the weight corresponding to the fraction of candidates in the full dataset at each score point. The metrics were calculated as follows:

$$MAE = \sum_{x=0}^{x_{max}} \frac{n_x}{N} * |y_x - Y_x|$$

$$bias = \sum_{x=0}^{x_{max}} \frac{n_x}{N} * (y_x - Y_x)$$

$$RMSE = \sqrt{\sum_{x=0}^{x_{max}} \frac{n_x}{N} * (y_x - Y_x)^2}$$

[15] Excluding the item used to split items into groups (this was not retained).

In all three equations, x is the score point on Form X, x_{max} is the maximum possible score on Form X, n_x is the number of candidates in the full dataset at score point x, N is the total number of candidates in the full dataset, y_x is the equated score for Form Y at that score point, and Y_x is the "true" equated score from the criterion equating.

Each accuracy metric gave one value *per method*, per *scenario, per test*. With 69 equating methods, 60 scenarios and 1,000 tests, that gives 4,140,000 accuracy values per metric. Accordingly, all interpretation has to be based on aggregate measures, most notably the mean and median, but also positions of quartiles. Therefore, we calculated these summary statistics and used boxplots to display the values. The accuracy metrics calculated were on the scale of marks, but as different tests had different numbers of marks, we converted these so that they can be presented as *percentages* of the Form Y maximum mark.

Although the metrics tell us slightly different things about equating method accuracy, we focus interpretation on results from MAE, as that is perhaps the simplest and most interpretable (i.e., how far away from the true equating relationship, on average, is that method?). Results relating to RMSE and bias will be addressed separately where they can add something new to our interpretation.

Approach to interpretation

The final thing to describe before presenting results is *how* we have interpreted outputs of the simulation. The first approach is, perhaps understandably, to take a 'big picture' view as to which methods perform best under different conditions, and which methods can be avoided. We therefore start by taking an overview of the results of all methods under the different scenarios, to try and answer these big, fundamental questions: this is, ultimately, the main purpose of the work.

The amount of data produced by the simulation also permits a much closer look at specific issues though. When trying carry out equating in practice, it may be necessary to choose between particular variants or methods, and even if there is broad, or high-level, guidance on which methods perform best, there might still be choices to make. Hence, using the vast amount of data, we can address some of these specific questions. After the 'big picture' results, we therefore address a series of specific questions about the methods, that seek to provide guidance around some of these smaller-scale issues.

Results: Outcomes of simulation

Effects of conditions on equating accuracy

The first result to examine is the impact of the different conditions on equating accuracy, as this provides the background against which all method comparisons are made. To examine these effects, we first focus on the results of the frequency estimation equipercentile method, with no smoothing, as this is arguably the most similar method to that used for criterion equating, and it also displays all the patterns to be discussed.

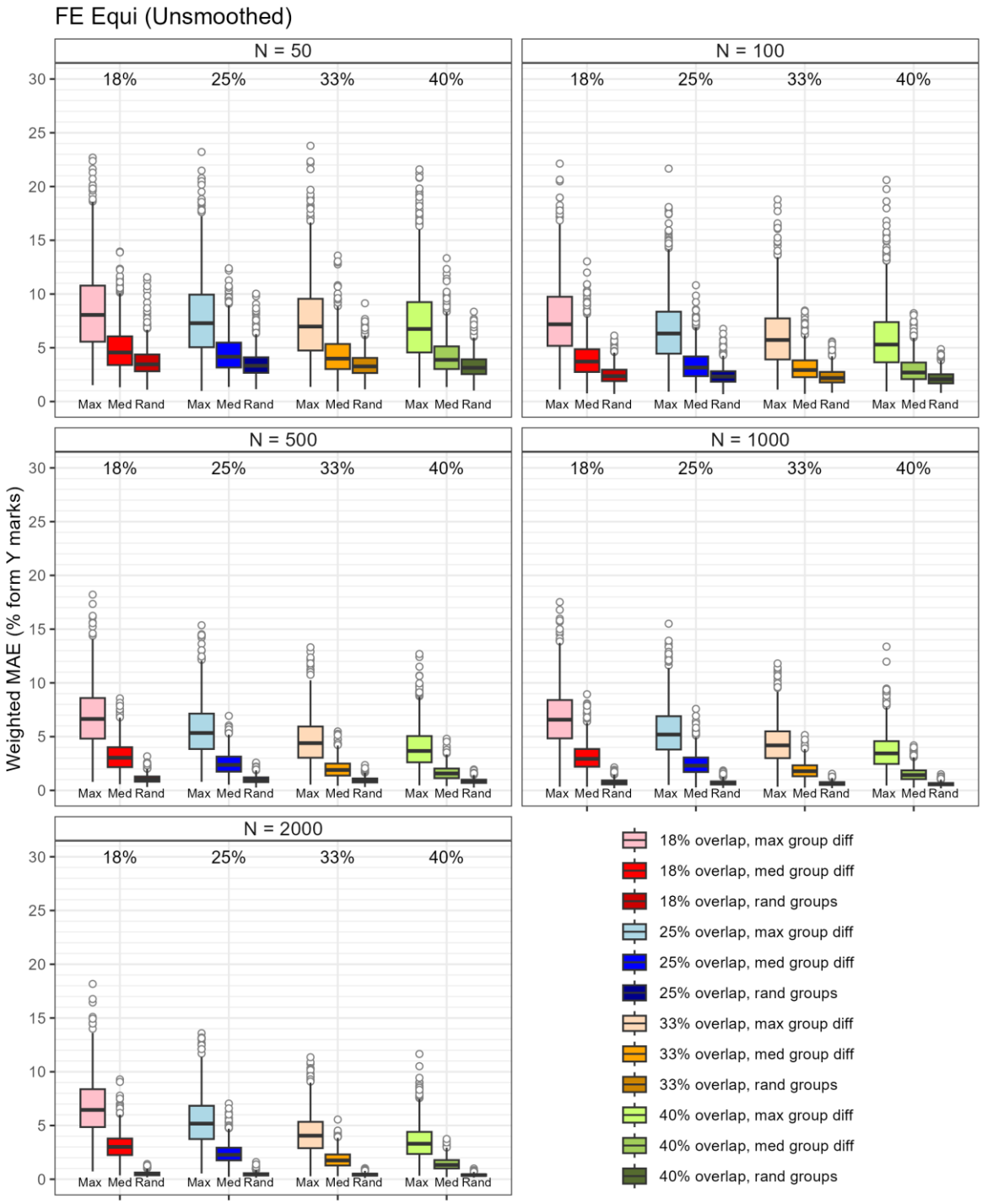

Figure 1. Weighted MAE (expressed as percentage of Form Y marks) for Frequency Estimation Equipercentile (unsmoothed) equating, under all simulation scenarios.

The three conditions (sample size, anchor length and group differences) provide predictable and consistent influences on equating accuracy. Moreover, the effects are in the direction we would expect *a priori*. Taking sample size first, we see that larger sample sizes are associated with smaller errors. In the N=50 panel (top-left), and looking at the maximum group difference boxes (the left box of each cluster), we see median weighted MAE values ranging from 8.1% to 6.8% of form Y marks. Conversely, in the N=2000 panel (bottom-left), the equivalent medians are smaller, ranging from 6.5% to 3.3%. The intermediate sample sizes of 100, 500, and 1,000 produce intermediate error values.

Looking next at group differences, we see that the largest errors are associated with the maximum group difference, and the smallest errors are associated with random groups. The differences are large in magnitude: looking at the N=50 (top-left panel) and 18% overlap condition (left-most cluster), the maximum group difference has a median weighted MAE of 8.1%, the medium group difference has a median of 4.6%, and the random grouping has a median of 3.5%. When N=2000 (bottom-left panel), the contrast is even greater, with the maximum group difference showing a median of 6.5%, and the random grouping showing a median of just 0.5%. Hence, the more similar the ability of the two populations of candidates, the smaller the error.

Finally, we see a smaller, but still clear, effect of anchor length, with longer anchors associated with smaller errors. With N=50 and maximum group difference, we see a median error of 8.1% for the 18% overlap condition, and 6.8% for the 40% overlap condition. The anchor length effect seems greater as sample sizes increase: with N=2000 and maximum group difference, we see a median error of 6.5% in the 18% overlap condition, but just 3.3% in the 40% overlap condition. The anchor length effect is much smaller in the random group conditions, such that in the N=2000 panel, it is almost imperceptible: the difference in median between the 18% and 40% overlap conditions is just 0.09 percentage points. In practice, this indicates that the need to have a long anchor test to facilitate equating is less important if the groups are more similar in ability.

While the simulation conditions produce similar effects in all equating methods, the magnitude of the effects varies between methods. To illustrate this, Figure 2 shows the same results for a different equating method (in this case, the method using a Partial Credit Model fitted via JML, with concurrent estimation). In contrast to Figure 1, the differences between conditions are small, often at ≤1 percentage point. In practice, this means that even though larger samples, longer anchors and smaller group differences should always lead to smaller equating errors, the extent to which this matters will vary between methods, with some highly sensitive to conditions, but others relatively robust: this is a theme that will be developed further in results discussed below. That is not, however, to say that a “robust” method will always be the best choice: a method that is highly sensitive to conditions might perform better than a “robust” alternative in some cases, and worse in others.

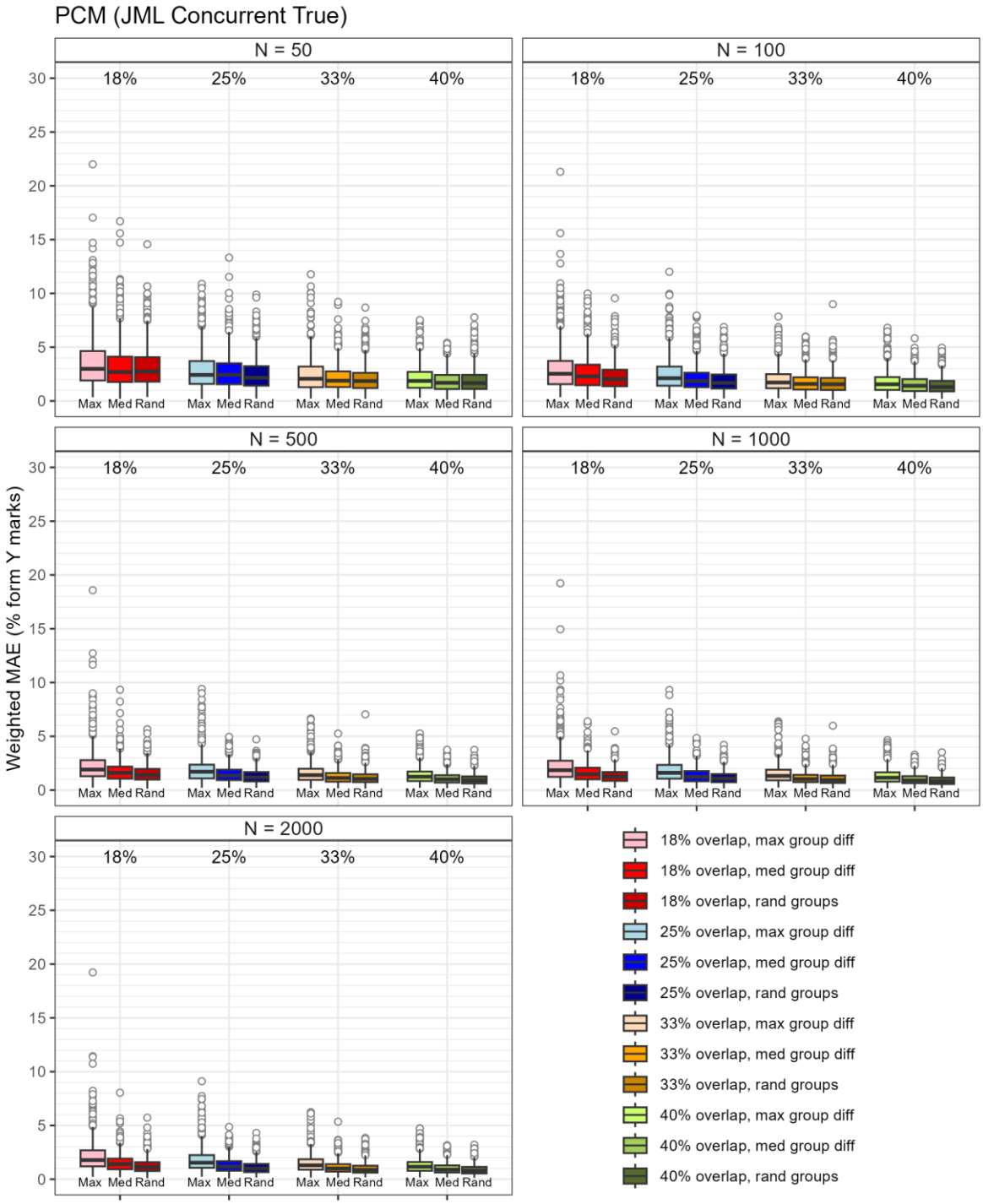

Figure 2. Weighted MAE (expressed as percentage of Form Y marks) for Partial Credit Model (JML concurrent True) equating, under all simulation scenarios.

Which equating methods are best?

Having established the effects of different simulation conditions, we can address the main focus of the analysis, which is to identify which equating methods perform best under the different conditions. Given that there are 60 scenarios, this potentially leads to 60 different answers, making interpretation challenging. However, as a first attempt at simplifying this problem, we can identify the method that produces the lowest error (here, the smallest mean weighted MAE) under each scenario: Table 4 shows these "winning" methods. Equivalent results derived from use of RMSE are almost identical; these are shown in Appendix 3.

Table 4. The best-performing equating methods (i.e., those with the smallest mean weighted MAE) under each simulation scenario.

Group difference	Form overlap	Sample size				
		50	**100**	**500**	**1000**	**2000**
Maximum	**18%**	Odds-transform (Nominal)	PCM (JML Conc True)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML SL Obs)
	25%	PCM (JML Conc True)	PCM (JML Conc True)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
	33%	PCM (JML Conc True)	PCM (JML Conc True)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML SL Obs)
	40%	Symm circle-arc (Levine)	PCM (JML Conc True)	GRM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
Medium	**18%**	Odds-transform (Chained SD)	Odds-transform (Chained SD)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
	25%	Odds-transform (Chained SD)	PCM (JML Conc True)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
	33%	Odds-transform (Chained SD)	PCM (JML Conc True)	GRM (MML Conc Obs)	GRM (MML Conc Obs)	GPCM (MML Conc Obs)
	40%	Odds-transform (Chained SD)	PCM (JML Conc True)	GRM (MML Conc Obs)	GRM (MML Conc Obs)	GPCM (MML Conc Obs)
Random	**18%**	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	25%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	33%	Odds-transform (Tucker SD)	Transformed Linear (Tucker)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	40%	Odds-transform (Tucker SD)	Transformed Linear (Tucker)	GRM (MML Conc Obs)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)

From this, some clear patterns emerge. First, only 10 methods are represented here, of the 69 trialled. This shows remarkable consistency in which methods performed best, with the majority of methods tested *never* "winning": this theme is picked up further below. Furthermore, just one method (GPCM, MML Concurrent Observed) is the "winner" in 18 of the 60 scenarios and a closely related method (GRM, MML Concurrent Observed) is the "winner" in a further 5; these methods together "won" in over a third of the scenarios.

Second, we see differing results across different sample sizes. Variants of odds-transform equating and PCM-based methods (specifically, using JML and concurrent estimation) dominate at N=50 and N=100, whilst GRM or GPCM-based methods (again using concurrent estimation) and frequency estimation equipercentile (with kernel smoothing) dominate once sample sizes are 500 or greater. There is also an effect of group differences. Frequency estimation equipercentile equating comes to dominate in random groups at large sample sizes, but GRM and GPCM-based methods dominate at the same sample sizes with medium or maximum group differences. Likewise, when N=100, odds-transform and transformed linear equating "win" with random groups, but PCM-based methods dominate in the other group conditions. Hence, both sample size and group difference strongly influence which equating method performs best.

Third, there is little apparent effect of anchor length on the "best" method, although there are some places where it may make a difference. Notably, when N=500 or N=1000, longer anchor lengths seem show better performance for GRM-based methods instead of GPCM-based methods (or, in one case, frequency estimation equipercentile equating). However, in no other scenarios do we see clear effects of anchor length.

Despite these methods being "best", this does *not* lead to the conclusion that all other methods should be disregarded. To explore this, Table 5 gives the mean (plus standard deviation) and median (plus 1st and 3rd quartiles) of weighted MAE scores for the *top 3* methods under each scenario (focusing just on 25% form overlap, given the limited effect of anchor length). From this, we see that the patterns derived from looking at the single "best" methods largely hold. For smaller samples, odds-transform or PCM (JML) methods generally dominate the top three. Likewise, for larger samples, GRM/GPCM methods are generally best, unless the groups are of equal ability, in which case variants of Frequency Estimation Equipercentile dominate.

There are, however, other findings of note, that show we should consider more than just the "best" methods. First, the differences in means between the top three methods are often extremely small, usually at ≤0.1 percentage points. This is particularly noticeable in larger sample sizes, where GRM and GPCM are almost interchangeable. Similarly, GPCM/GRM models fitted to both tests concurrently dominate, but separate model fitting and use of the Stocking-Lord method performs similarly. As another example, in the random group scenarios where Frequency Estimation Equipercentile equating dominates, kernel smoothing comes out best, but other smoothing methods have means only a fraction of a percentage point away from the "best" result. Hence, the magnitude of differences between the "best" method and the second, or third (or, indeed, fourth or fifth in some cases) may be very small.

Second, Table 5 shows that the method of evaluation could change our interpretation of which method is best. If the *median* error were used to rank performance, most results would remain the same, but some would change. For example, when N=100 and with a medium group difference, the odds-transform (chained) method and PCM (MML, observed scores) would "overtake" the PCM (JML) if median values were used. Note, though, that the differences between these methods only really occur at the *third* decimal place, showing how similar they are. Hence, this emphasises that not all methods outside the "best" can be readily discarded.

Table 5. The top three methods for each scenario under the 25% form overlap condition.

Sample size	Group diff.	Rank	Method	Mean	SD	Median	Q1	Q3
50	Max	1	PCM (JML Concurrent True)	2.86	1.79	2.42	1.58	3.71
		2	Symmetric circle-arc (Nominal)	2.92	1.82	2.53	1.59	3.90
		3	Odds-transform (Levine)	2.92	1.82	2.51	1.64	3.66
	Med	1	Odds-transform (Chained SD)	2.55	1.38	2.33	1.52	3.31
		2	Symmetric circle-arc (Levine)	2.62	1.58	2.29	1.46	3.42
		3	Odds-transform (Levine)	2.62	1.58	2.29	1.47	3.43
	Rand	1	Odds-transform (Tucker SD)	2.08	1.15	1.87	1.25	2.70
		2	Odds-transform (Chained SD)	2.18	1.21	1.93	1.29	2.82
		3	Symmetric circle-arc (Tucker)	2.20	1.22	1.98	1.31	2.83
100	Max	1	PCM (JML Concurrent True)	2.47	1.52	2.12	1.41	3.20
		2	Symmetric circle-arc (Levine)	2.51	1.59	2.14	1.39	3.27
		3	Odds-transform (Levine)	2.51	1.59	2.14	1.39	3.28
	Med	1	PCM (JML Concurrent True)	2.07	1.13	1.85	1.28	2.63
		2	Odds-transform (Chained SD)	2.08	1.18	1.84	1.20	2.69
		3	PCM (MML Concurrent Obs)	2.12	1.28	1.84	1.23	2.74
	Rand	1	Odds-transform (Tucker SD)	1.65	0.89	1.47	0.98	2.13
		2	Transformed Linear (Tucker)	1.66	0.85	1.52	1.04	2.15
		3	Odds-transform (Chained SD)	1.71	0.91	1.56	1.03	2.21
500	Max	1	GPCM (MML Concurrent Obs)	1.73	1.39	1.37	0.92	2.11
		2	GPCM (MML SL Obs)	1.78	1.34	1.44	0.93	2.19
		3	GRM (MML SL Obs)	1.81	1.38	1.49	0.95	2.22
	Med	1	GPCM (MML Concurrent Obs)	1.18	0.75	1.02	0.70	1.49
		2	GRM (MML Concurrent Obs)	1.21	1.24	1.01	0.71	1.47
		3	GPCM (MML SL Obs)	1.24	0.77	1.07	0.70	1.57
	Rand	1	FE Equi (Kernel smoothed)	0.87	0.37	0.80	0.61	1.07
		2	Chained Equi (Kernel smoothed)	0.98	0.40	0.91	0.69	1.19
		3	FE Equi (Loglin smoothed)	0.99	0.36	0.92	0.75	1.18
1000	Max	1	GPCM (MML Concurrent Obs)	1.51	1.24	1.18	0.77	1.85
		2	GPCM (MML SL Obs)	1.53	1.17	1.25	0.81	1.83
		3	GRM (MML SL Obs)	1.58	1.19	1.28	0.84	1.92
	Med	1	GPCM (MML Concurrent Obs)	0.92	0.59	0.79	0.53	1.12
		2	GPCM (MML SL Obs)	0.95	0.59	0.82	0.56	1.20
		3	GRM (MML SL Obs)	0.96	0.62	0.82	0.56	1.21
	Rand	1	FE Equi (Kernel smoothed)	0.64	0.26	0.59	0.46	0.77
		2	FE Equi (Loglin smoothed)	0.70	0.25	0.66	0.52	0.83
		3	FE Equi (Bump smoothed)	0.71	0.25	0.67	0.54	0.83
2000	Max	1	GPCM (MML Concurrent Obs)	1.40	1.21	1.12	0.72	1.61
		2	GPCM (MML SL Obs)	1.41	1.14	1.10	0.74	1.67
		3	GRM (MML SL Obs)	1.45	1.16	1.14	0.76	1.71
	Med	1	GPCM (MML Concurrent Obs)	0.81	0.53	0.70	0.48	1.01
		2	GRM (MML Concurrent Obs)	0.84	0.83	0.71	0.48	1.01
		3	GPCM (MML SL Obs)	0.85	0.55	0.73	0.49	1.04
	Rand	1	FE Equi (Kernel smoothed)	0.44	0.18	0.42	0.32	0.53
		2	FE Equi (Loglin smoothed)	0.47	0.18	0.44	0.34	0.56
		3	FE Equi (Bump smoothed)	0.47	0.18	0.45	0.35	0.56

From these first analyses, we can already draw some conclusions. At smaller sample sizes, odds-transform equating and PCM-based methods (using JML) generally perform best. If groups are truly random and sample sizes are small, transformed linear equating also performs well. At larger sample sizes, equating methods based on fitting the GRM or GPCM to data (ideally fitting the model to both tests concurrently) work best, unless the groups are truly random, in which case Frequency Estimation Equipercentile equating works well.

Given the dominance of methods based on Rasch or IRT across many conditions in Table 5, we also identified the best-performing equating methods based only on *classical* techniques, as there may be circumstances where Rasch or IRT cannot be used. The resulting "best" classical methods are shown in Table 6.

From Table 6, we see strong, and sometimes somewhat surprising, patterns of which methods perform best. First, in the maximum group difference condition, symmetric circle-arc equating using the Levine approach to estimating group means "wins" in *virtually every condition*. This is arguably surprising, as circle-arc methods are considered to be useful for small sample sizes, but here the method even comes out best at the largest sample sizes tested when there is a large group difference. It is important to note, however, that although symmetric circle-arc "wins" in each of these conditions, odds-transform equating with the Levine approach to estimating group means gives almost identical results in each case. This can be seen in Table 5, in the maximum group difference conditions for N=50 and N=100. In these cases, symmetric circle-arc and odds transform methods produce mean weighted MAE values that are the same to two decimal places. Furthermore, when N=50, the *median* weighted MAE for odds-transform equating is actually *smaller* than that for symmetric circle-arc. Hence, although Table 6 shows symmetric circle-arc to perform well, odds-transform equating performs almost identically. This comparison is explored further below.

From Table 6, we can also see how for N=50 and N=100, with medium group differences or random groups, odds-transform variants dominate. With medium group differences, the chained variant "wins" in all conditions, whilst for random groups, the Tucker variant "wins" in most conditions (with transformed linear equating "winning" with long anchors). In these conditions, differences between odds-transform and symmetric circle-arc are still small, but typically occur in the first decimal place.

Finally, Table 6 also shows how frequency estimation equipercentile methods dominate in large sample sizes for medium group differences and random groups. For N=500, N=1000 and N=2000, the modified Wang-Brennan unsmoothed variant "wins" almost every condition with medium group differences, whilst the kernel smoothed variant "wins" every condition with random groups. Hence, if GPCM/GRM methods are unable to be used with large sample sizes and small-to-medium group ability differences, frequency estimation equipercentile methods appear to be the best alternative.

Table 6. The best-performing classical equating methods under each simulation scenario (i.e., those with the smallest mean weighted MAE once Rasch/IRT methods are excluded).

Group difference	Form overlap	Sample size				
		50	100	500	1000	2000
Maximum	18%	Odds-transform (Nominal)	Odds-transform (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)
	25%	Symmetric circle-arc (Nominal)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)
	33%	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)
	40%	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	FE Equi (Modified WB Unsmoothed)
Medium	18%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	Linear (Levine True)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)
	25%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)
	33%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)
	40%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	Chained Equi (Kernel smoothed)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)
Random	18%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	25%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	33%	Odds-transform (Tucker SD)	Transformed Linear (Tucker)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	40%	Odds-transform (Tucker SD)	Transformed Linear (Tucker)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)

Thus far, we have focused on the “best” methods. However, we should also consider which methods worked acceptably, and which did not work well at all. To introduce this, we present a series of figures showing results from *all* methods trialled in the simulation. To reduce the number of scenarios considered, we again focus on the 25% overlap condition, and on the medium group difference condition. These represent a set of scenarios that could make equating somewhat challenging, but which might also be encountered in practice. Figure 3 shows results for N=50 and Figure 7 shows results for N=2000: figures between correspond to intermediate sample sizes. Equivalent figures for random groups and maximum group differences are shown in Appendix 3.

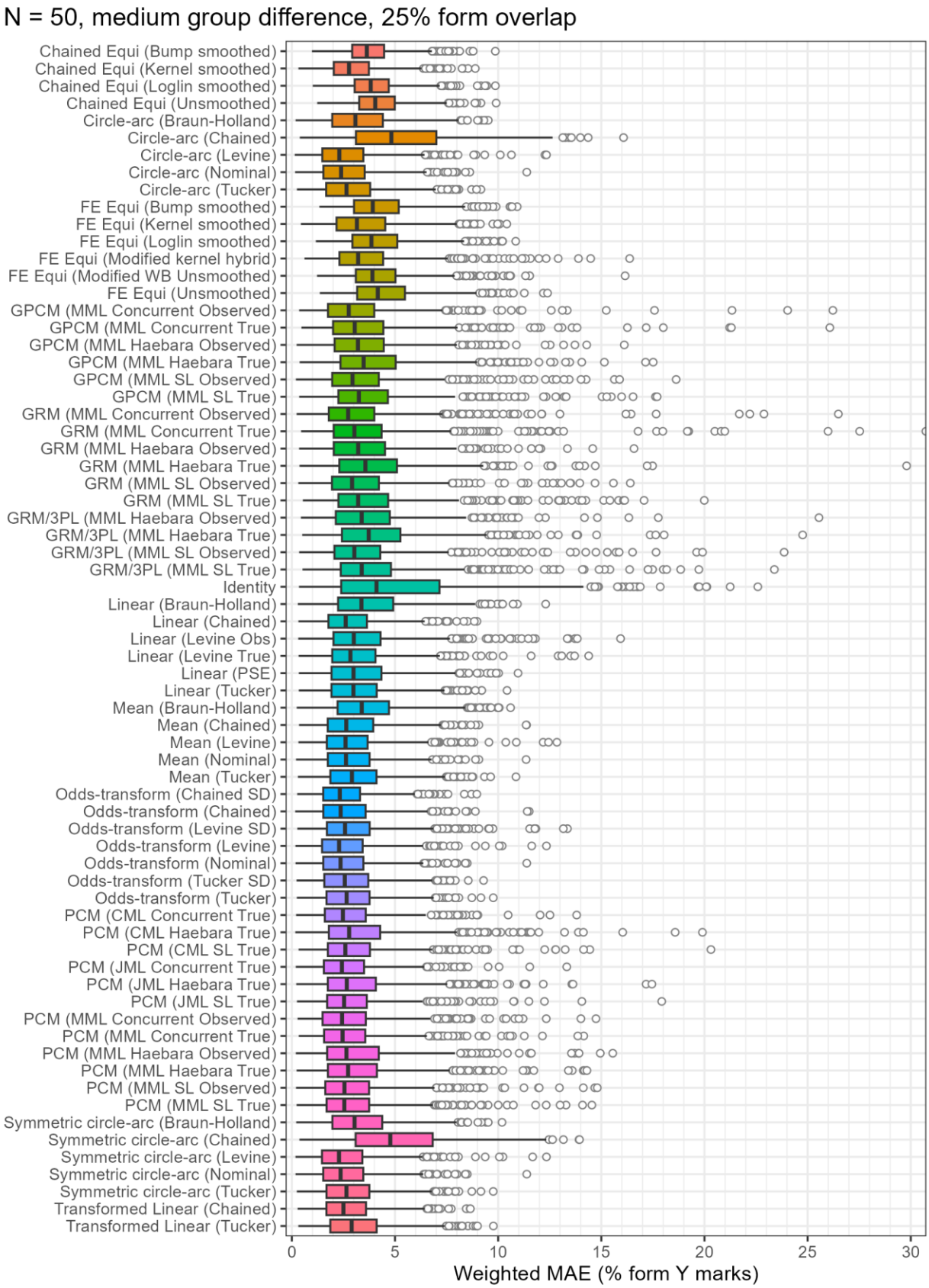

Figure 3. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 50, 25% form overlap and a medium difference between groups. Note that outliers with values >30% are not shown.

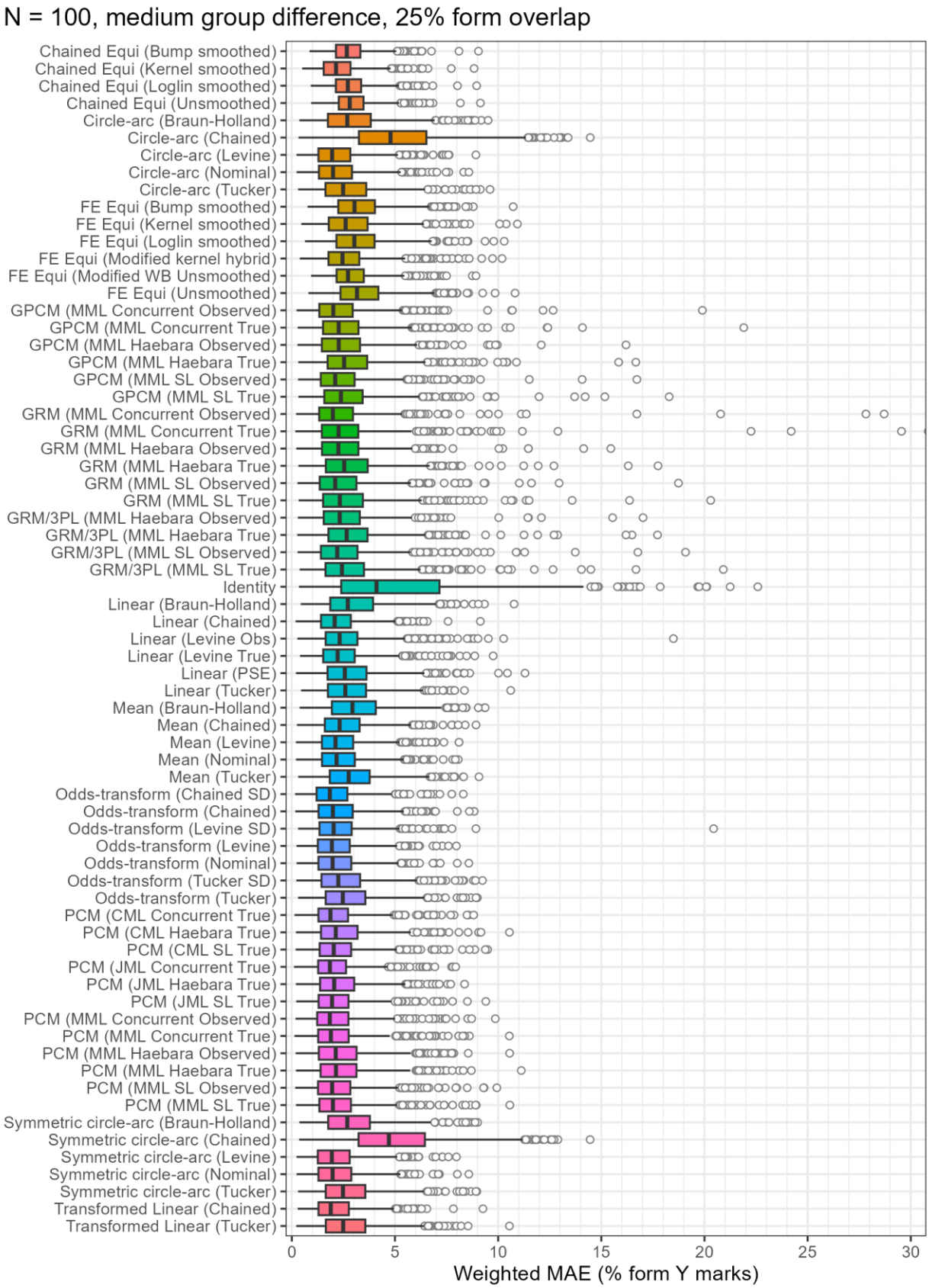

Figure 4. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 100, 25% form overlap and a medium difference between groups. Note that outliers with values >30% are not shown.

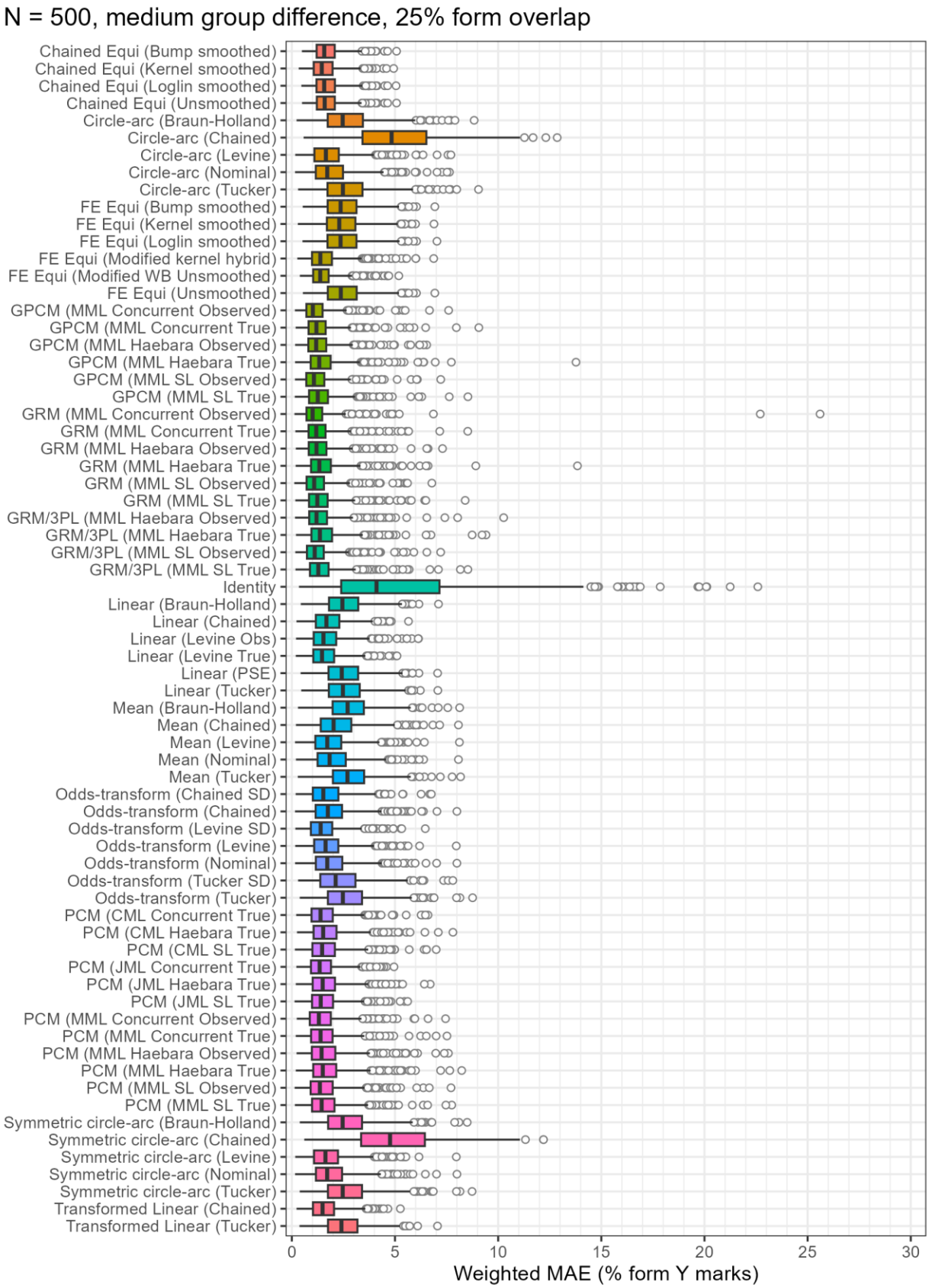

Figure 5. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 500, 25% form overlap and a medium difference between groups. Note that outliers with values >30% are not shown.

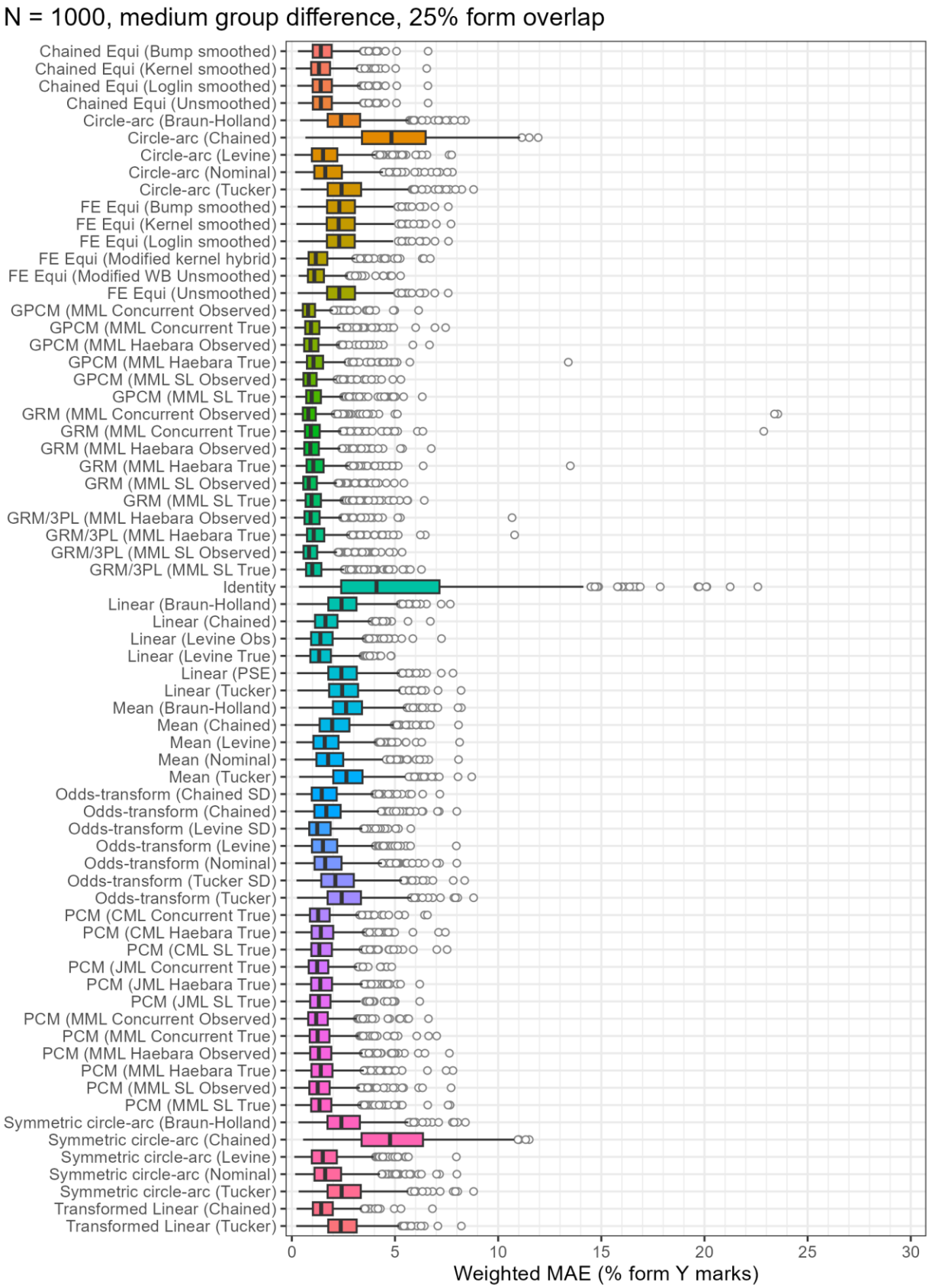

Figure 6. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 1,000, 25% form overlap and a medium difference between groups. Note that outliers with values >30% are not shown.

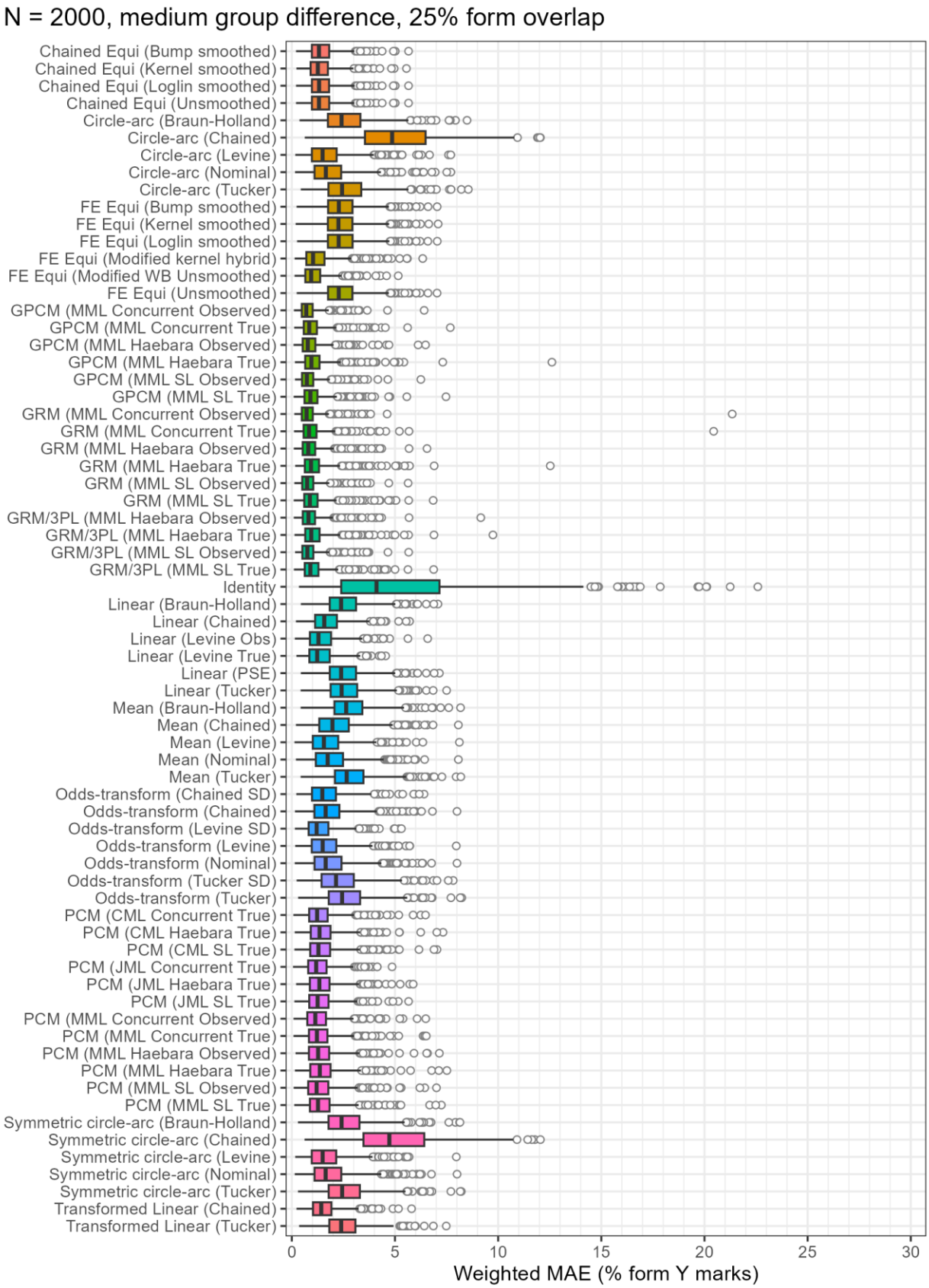

Figure 7. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 2,000, 25% form overlap and a medium difference between groups. Note that outliers with values >30% are not shown.

From these figures, the first thing to note is that many methods show somewhat similar results. Although a few methods do show clearly poorer performance (which will be discussed below), *most* seem broadly comparable. This is reassuring: if equating has been carried out using methods other than the "best" ones, the results should still be acceptable. However, the broadly similar performance might mask patterns where particular methods are consistently outperformed by alternatives: this will also be discussed below.

Some methods show consistently *reasonable* performance. That is, they may be "best" only under certain conditions, if at all, but under most conditions they are acceptable. This is most evident for PCM (most variants, but particularly those based on JML estimation), odds-transform equating (particularly the Levine variant) and symmetric circle-arc equating (again, particularly the Levine variant). As discussed above, these methods perform best at small sample sizes, but they also perform adequately at larger sample sizes; this is particularly evident under the maximum group difference condition (see Appendix 3), when other methods (e.g. Frequency Estimation Equipercentile) perform substantially *worse*, but PCM, odds-transform (Levine) and symmetric circle-arc (Levine) perform consistently *adequately*.

As noted above, most methods show decreasing error with increasing sample size. However, one method notably does *not* show that pattern: identity equating shows the same error regardless of conditions. Hence, at N=50, it might *appear* to be broadly comparable to other methods, but by N=500 it is performing worse than other methods, and by N=2000 it is substantially worse[16]. Chained circle-arc equating and mean equating methods show a similar pattern, albeit those methods do show *some* improvement as sample sizes increase. Nevertheless, these methods might, at the smallest sample sizes, appear reasonable, but there would always be better options, especially as sample sizes increase. Care should be taken, then, to consider performance under different conditions: some methods only appear reasonable under specific conditions. An examination of methods that are *always* outperformed by other methods is described below.

Another pattern to note is the presence of outliers, most notably for IRT-based methods. Although GRM and GPCM performed well at large sample sizes, they produced *some* very poor results at sample sizes of 50 and 100, suggesting caution would have to be applied if using them with such small samples. We must remember, of course, that the majority of cases produced a degree of accuracy in keeping with other methods. However, it shows a lack of reliability at these sample sizes, which is probably due to numerical issues such as lack of algorithmic convergence in model fitting. At sample sizes of 500 or over, outliers produced by these methods were similar to those produced by any other method, while their medians and inter-quartile ranges became *smaller* than other methods, showing better performance. However, if choosing methods to apply in real settings, and consistency of outcomes is important, occurrence of outliers like this might need to be taken into account.

The charts and tables in this section have focussed upon weighted mean absolute error (MAE). We have not included extensive discussion on other metrics, in particular bias. Therefore, to briefly illustrate what results indicated about bias, in Figure 8 we show a single plot, for N=2000 and maximum group difference, with 25% form overlap.

[16] Naturally, if we had specific external reasons to expect test forms to be of similar difficulty (perhaps from small scale pretesting or expert judgement) then identity equating might perform better.

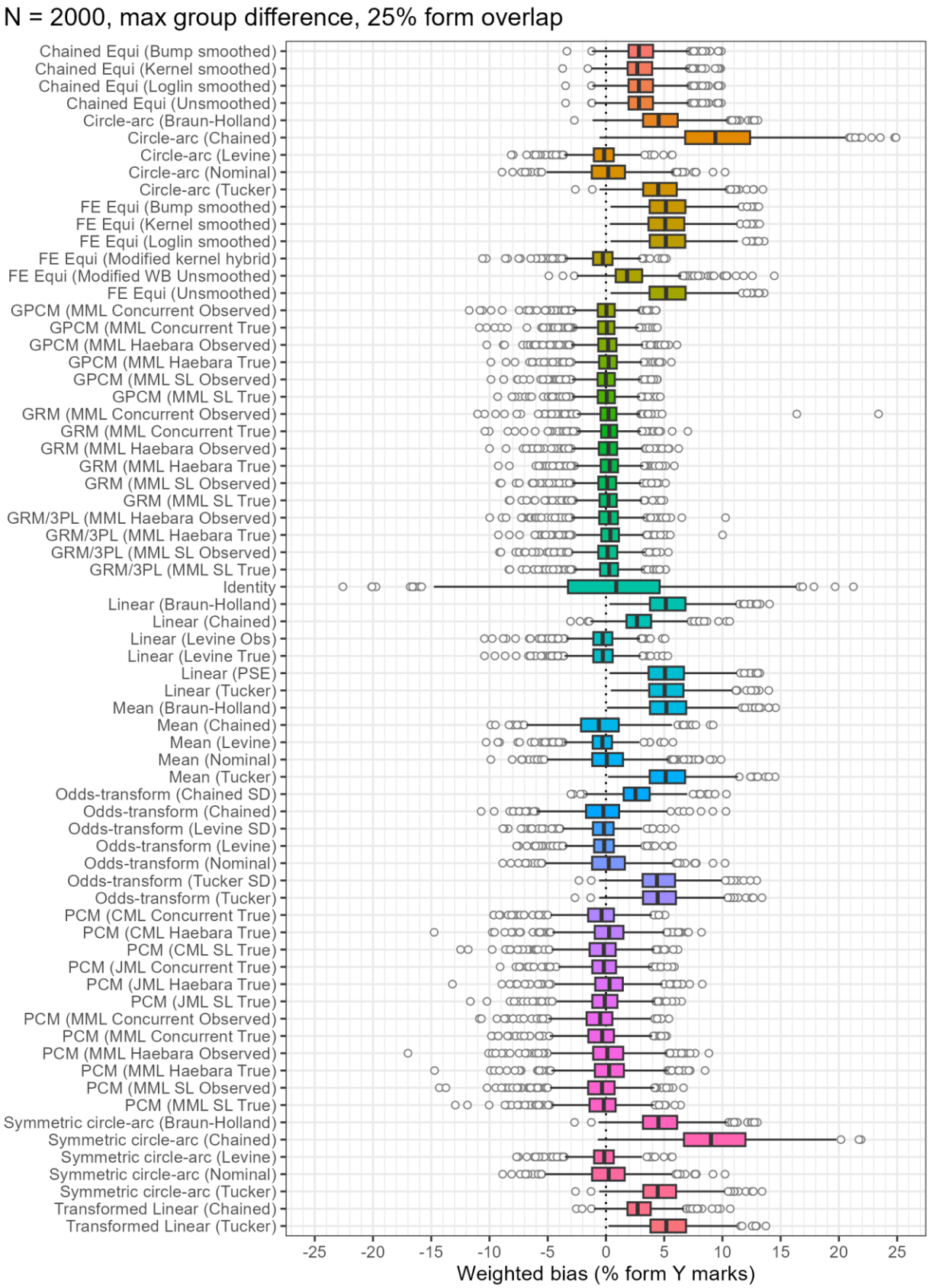

Figure 8. Box plots of weighted bias values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 2,000, 25% form overlap and a large difference between groups. Note that outliers with values >25% (or < -25%) are not shown.

As in previous research studies (for example, Sinharay & Holland, 2007), Figure 8 shows that many classical methods display significant bias when there are large ability differences between groups. The methods may not fully adjust for the differences between groups (i.e., candidates taking Form Y had higher ability) meaning that the difficulty of Form Y was underestimated (i.e., scores on Form X were equated to scores on Form Y that were too high). This issue is particularly evident for all forms of chained equating, all forms of circle-arc, linear (including transformed linear) or mean equating *other than the Levine or nominal variants*, and all forms of frequency estimation equipercentile equating *other than the modified Wang-Brennan and modified kernel hybrid variants*. The biases of chained equipercentile methods were noticeably smaller than that of equivalent frequency estimation (apart from modified Wang-Brennan and modified kernel hybrid variants) but still significant. Similarly, chained linear equating was less biased than linear equating using the Tucker method. Hence, in these classical methods, approaches that are explicitly designed to account for measurement error in the anchor do appear to display lower levels of bias in these situations. We note also that methods based upon IRT or Rasch PCM models displayed much lower level of bias. Overall, these results help to explain why equating based upon frequency estimation or the Tucker method only tended to "win" in scenarios with random differences between the groups taking different test forms.

Which equating methods can we dispose of?

Having explored which methods are most effective, we now attempt to better identify those methods that are best avoided. Specifically, we look for instances where one method is *always* outperformed by another specific method across every scenario. By a "scenario" we mean one of the 60 combinations of conditions (5 sample sizes by 4 anchor lengths by 3 levels of group differences) that we have considered. We calculate the mean of the weighted mean absolute error (as a percentage of the test maximum) of each method in each scenario. If one method has a higher mean weighted MAE than another in a given scenario we might say it is outperformed in this scenario. If one method is outperformed by another specific method across every one of the 60 scenarios, we say that the one method is *dominated* by the other. If a method is dominated, this implies that we would be better using a different method all the time regardless of the scenario we are in. Conversely, for any two methods that are never dominated by each other, there is always at least one scenario where one method is preferable to the other.

Table 7 shows the 47 methods that are dominated by another method. For example, looking at the first row, unsmoothed frequency estimation equipercentile equating is dominated by frequency estimation equipercentile equating with loglinear smoothing. This implies that for every scenario considered in this study, the technique based on loglinear smoothing outperforms the unsmoothed method. As such, we can infer that, for sample sizes of 2,000 or below, we are always better off if we apply some form of smoothing in equipercentile equating, even if the resulting difference is not large (as shown above).

Table 7. Summary of methods dominated by other methods (i.e., methods outperformed by other methods in *every* simulated scenario).

Dominated method	N methods dominating	Dominated By
Chained Equi (Unsmoothed)	3	Chained Equi (Bump smoothed), Chained Equi (Loglin smoothed), Chained Equi (Kernel smoothed)
Chained Equi (Bump smoothed)	1	Chained Equi (Kernel smoothed)
Chained Equi (Loglin smoothed)	1	Chained Equi (Kernel smoothed)
Circle-arc (Chained)	12	12 methods
Circle-arc (Nominal)	2	Odds-transform (Nominal), Symmetric circle-arc (Nominal)
Circle-arc (Tucker)	3	Odds-transform (Tucker), Odds-transform (Tucker SD), Symmetric circle-arc (Tucker)
Circle-arc (Levine)	2	Odds-transform (Levine), Symmetric circle-arc (Levine)
Circle-arc (Braun-Holland)	2	Odds-transform (Tucker SD), Symmetric circle-arc (Braun-Holland)
FE Equi (Unsmoothed)	2	FE Equi (Loglin smoothed), FE Equi (Kernel smoothed)
FE Equi (Bump smoothed)	1	FE Equi (Kernel smoothed)
FE Equi (Loglin smoothed)	1	FE Equi (Kernel smoothed)
GPCM (MML SL True)	6	GPCM (MML SL Observed), GPCM (MML Concurrent True), GPCM (MML Concurrent Observed), GRM/3PL (MML SL Observed), GRM (MML SL Observed), GRM (MML Concurrent Observed)
GPCM (MML Haebara True)	11	11 methods
GPCM (MML Haebara Observed)	3	GPCM (MML SL Observed), GPCM (MML Concurrent Observed), GRM (MML SL Observed)
GPCM (MML Concurrent True)	2	GPCM (MML Concurrent Observed), GRM (MML Concurrent Observed)
GRM/3PL (MML SL True)	9	9 methods
GRM/3PL (MML SL Observed)	1	GPCM (MML Concurrent Observed)
GRM/3PL (MML Haebara True)	12	12 methods
GRM/3PL (MML Haebara Observed)	6	GPCM (MML SL Observed), GPCM (MML Concurrent Observed), GRM/3PL (MML SL Observed), GRM (MML SL Observed), GRM (MML Haebara Observed), GRM (MML Concurrent Observed)
GRM (MML SL True)	6	GPCM (MML SL Observed), GPCM (MML Concurrent True), GPCM (MML Concurrent Observed), GRM/3PL (MML SL Observed), GRM (MML SL Observed), GRM (MML Concurrent Observed)
GRM (MML Haebara True)	11	11 methods
GRM (MML Haebara Observed)	2	GPCM (MML Concurrent Observed), GRM (MML SL Observed)
GRM (MML Concurrent True)	4	GPCM (MML SL Observed), GPCM (MML Concurrent Observed), GRM (MML SL Observed), GRM (MML Concurrent Observed)
Identity	38	38 methods
Linear (Levine Observed)	2	GPCM (MML Concurrent Observed), GRM (MML Concurrent Observed)
Linear (Levine True)	1	PCM (MML Concurrent Observed)
Linear (Chained)	1	Transformed Linear (Chained)
Linear (PSE)	1	Transformed Linear (Chained)
Mean (Chained)	14	14 methods
Mean (Nominal)	10	10 methods

Dominated method	N methods dominating	Dominated By
Mean (Tucker)	5	Circle-arc (Tucker), Odds-transform (Chained SD), Odds-transform (Tucker), Odds-transform (Tucker SD), Symmetric circle-arc (Tucker)
Mean (Levine)	7	Circle-arc (Levine), Odds-transform (Levine), PCM (JML Concurrent True), PCM (MML Concurrent True), PCM (MML Concurrent Observed), PCM (CML Concurrent True), Symmetric circle-arc (Levine)
Mean (Braun-Holland)	9	9 methods
Odds-transform (Chained)	2	Odds-transform (Levine), Symmetric circle-arc (Levine)
Odds-transform (Levine SD)	2	PCM (JML Concurrent True), PCM (MML Concurrent Observed)
PCM (MML SL True)	4	PCM (JML SL True), PCM (JML Concurrent True), PCM (MML Concurrent True), PCM (CML Concurrent True)
PCM (JML SL True)	1	PCM (JML Concurrent True)
PCM (JML Haebara True)	2	PCM (JML SL True), PCM (JML Concurrent True)
PCM (MML SL Observed)	1	PCM (MML Concurrent Observed)
PCM (MML Haebara True)	9	9 methods
PCM (MML Haebara Observed)	3	PCM (JML Concurrent True), PCM (MML SL Observed), PCM (MML Concurrent Observed)
PCM (MML Concurrent True)	1	PCM (JML Concurrent True)
PCM (CML SL True)	4	PCM (JML SL True), PCM (JML Concurrent True), PCM (MML Concurrent True), PCM (CML Concurrent True)
PCM (CML Haebara True)	9	9 methods
PCM (CML Concurrent True)	1	PCM (JML Concurrent True)
Symmetric circle-arc (Chained)	9	9 methods
Symmetric circle-arc (Braun-Holland)	1	Odds-transform (Tucker SD)

To summarise the findings in Table 7:

- Unsmoothed equipercentile equating is dominated by smoothed methods. Kernel smoothing dominates every other smoothing approach.
- Both Stocking-Lord (SL) and Haebara approaches to Rasch PCM equating are dominated by approaches based on concurrent analysis. Furthermore, the JML approach to analysis dominates the CML approach. Having said this, we note that different R packages were used for different approaches so the results may reflect differences in technicalities (e.g., convergence criteria) rather than superior performance of a method *per se*.
- For other IRT methods, Haebara approaches are dominated by both the Stocking-Lord (SL) method and by concurrent analysis. Observed score approaches dominate true score approaches. Models including the 3PL are dominated by those that avoid this.
- Mean equating is always dominated by an alternative based on circle-arc equating.
- Circle-arc equating (i.e., non-symmetric circle-arc equating) is always dominated by equivalent approaches based upon symmetric circle-arc equating or odds-transform equating.
- Many approaches (even including mean equating) dominate identity equating, meaning this approach should never be adopted for the sample sizes explored in this study.

If we exclude the approaches that are dominated by another method (Table 7), this leaves a much smaller subset of methods for us to consider working with. From a set of 69 methods that

we started with (see Table 1), we are left with just 22 that are worth considering for use in practice: these are listed in Table 8. This means that for every method listed in Table 8, if it is compared to any other single method, there is at least one scenario out of the 60 considered where it is superior. Having said this, further analysis later in this report will reveal that even some of these 22 methods are *also* not worth retaining. This might be because its superiority over another listed method in Table 1 is so rare and so small it is not worth considering.

Table 8. Methods not dominated by any others in the main simulation run.

IRT or classical	Type	N variants	Variants
Classical	FE Equi	3	Modified WB Unsmoothed, Modified kernel hybrid, Kernel smoothed
Classical	Chained Equi	1	Kernel smoothed
Classical	Linear	2	Braun-Holland, Tucker
IRT	PCM	2	MML Concurrent Observed, JML Concurrent True
IRT	GRM	2	MML SL Observed, MML Concurrent Observed
IRT	GPCM	2	MML SL Observed, MML Concurrent Observed
Classical	Symmetric circle-arc	3	Nominal, Levine, Tucker
Classical	Odds-transform	5	Chained SD, Levine, Tucker, Tucker SD, Nominal
Classical	Transformed Linear	2	Chained, Tucker

Sensitivity to non-experimentally controlled factors

The above results are based upon averaging error across all 1,000 datasets within each scenario. However, it is also possible to rerun analysis on subsets of data that happened to display particular characteristics. For example, we might want to focus upon instances where the random separation of items happened to result in two test forms with very different levels of difficulty. A small sensitivity analysis was undertaken to explore these issues.

Firstly, we explored how results changed if we restricted to instances where the test forms being equated had substantially different difficulty. This was judged by the extent to which the criterion equating was different from identity equating. Results of showed that this restriction had hardly any impact upon results. This implies that the results we have described above are not reliant upon the test forms having similar levels of difficulty.

We also explored how the pattern of results changed dependent upon the overall maximum score of each form and also the related matter of the reliability of the anchor test. The analysis showed that this factor did have some impact upon results. In particular, in small samples (100 or less), IRT methods based upon MML estimation tended to have lower error if the overall maximum score of each test form was higher. However, given that we would not generally recommend such methods for use in small samples in any case, this has no impact upon the major conclusions of our analysis.

Finally, we explored whether our results change if, rather than focussing on the weighted absolute error across the whole score range, we focussed on accuracy at particular points (as might be important in grade boundaries or pass marks, for example). Specifically, we looked at absolute equating error at form X scores representing 20 and 80 per cent of the maximum available. This made some noticeable differences to results, particularly at the 20 per cent point. This is likely to be because, for most of the tests included in the study, this would represent an unusually low level of performance, so there is little available data at this end of the score range. The impact of this is that, for small samples, methods such as equipercentile equating perform worse. Separately, due to their inability to capture non-linear relationships at the extremes of the score range, methods such as mean equating and linear equating also perform worse.

Overall, the sensitivity analysis does not alter the main conclusions of our analysis or our recommendations about the most appropriate equating techniques. However, reviewing this analysis does reiterate the importance, in practice, of checking the stability of equating by calculating standard errors of equating *at the score points at which they will be used*.

Analysis later in this section will explore the impact of IRT model fit on results.

Secondary questions about IRT equating

As noted above, now that the ‘big-picture’ questions have been addressed, we can also address more specific issues that can, and do, arise when choosing equating methods. In particular, above, we found that PCM (fitted by JML) and GPCM or GRM-based methods outperformed all others in most conditions. Therefore, this section explores questions relating to different ways of implementing IRT equating. All of these are looked at in a pairwise manner to address whether to do one thing or the other when faced with two approaches that, on the face of things, are very similar and equally valid.

It is important to emphasize that, for the most part, the differences found between different (but very similar) techniques were tiny. As such, the results here cannot be used to dismiss the less successful approaches as inadequate. Nonetheless, analysts often have to make decisions about whether to choose one approach over another and justify the choices they have made. Therefore, even though the differences in performance are tiny, the results are often consistent enough to provide a clear answer as to which method is generally preferable.

GPCM vs GRM

We begin with a comparison where no visible difference in either direction was identified – the comparison between using the GRM or the GPCM to capture the fact that different items may have different levels of discrimination. We focus on the results for observed score equating as, for these methods, observed score equating was found to be more accurate (see later section). For the time being we also focus upon results from the Stocking-Lord approach (a comparison to other approaches will be provided later).

Figure 9 shows the distribution of differences in the weighted mean absolute error (MAE) of equating across the 1,000 analyses conducted in each of 60 separate scenarios. The 60 scenarios are distinguished by the extent of overlap between test forms (4 conditions in the columns of the panel), the size of differences between candidate groups (3 conditions in the panel rows), and sample size (5 conditions). Each point denotes the mean difference in weighted MAE between the GRM and GPCM approach and the error bars denote the range between the 90th and 10th percentiles of differences. Points below zero indicate that the GPCM approach was superior whereas points above zero indicate that the GRM was superior.

As can be seen, the equating error of the approaches was always very close. For sample sizes of 500 or more, the weighted MAEs of the two methods were within 0.5 per cent of each other at least 80 per cent of the time (i.e., the error bars are within the range -0.5 to +0.5). Furthermore, there is no visible tendency for either technique to outperform the other. As such, there is no empirical reason to prefer the use of one over the other.

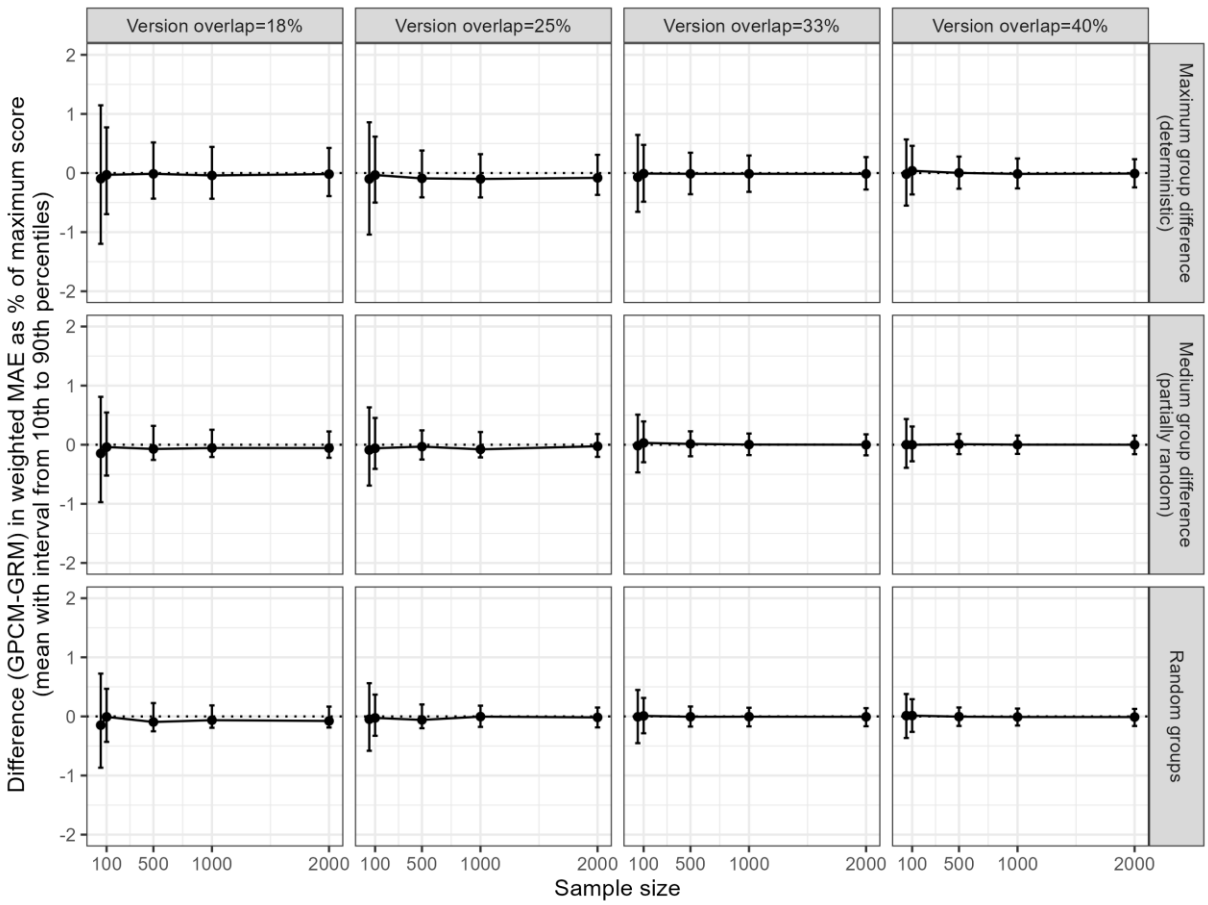

Figure 9. The distribution of differences in weighted MAE between GPCM and GRM approaches to observed score equating across a range of scenarios.

Rasch vs more complex IRT and True vs Observed score equating

Next, we compare the performance of using a Rasch model (specifically, a PCM) to that using a more complex approach based on the GRM. We also compare the observed score and true score approaches to equating. For consistency, we focus on models estimated using MML[17]. We do not compare to the GPCM as the previous section has already established that this performs extremely similarly to the GRM.

The boxplots in Figure 10 show the distribution of weighted MAE for each of four methods (GRM or Rasch by True or Observed score equating) across 27 scenarios[18]. For the GRM (red boxes), in every scenario, an observed score approach (solid lines) tends to have lower error than a true score approach (dashed lines). In contrast, for the PCM (blue boxes), there is very little difference in performance between true and observed score equating.

If the extent of overlapping items between the two test forms is small (i.e., test form overlap of 18%) then, for sample sizes below 500, Rasch equating tends to have lower error than observed score equating based on the GRM. However, where we have a greater degree of overlap, the GRM approach can be just as accurate even at small sample sizes.

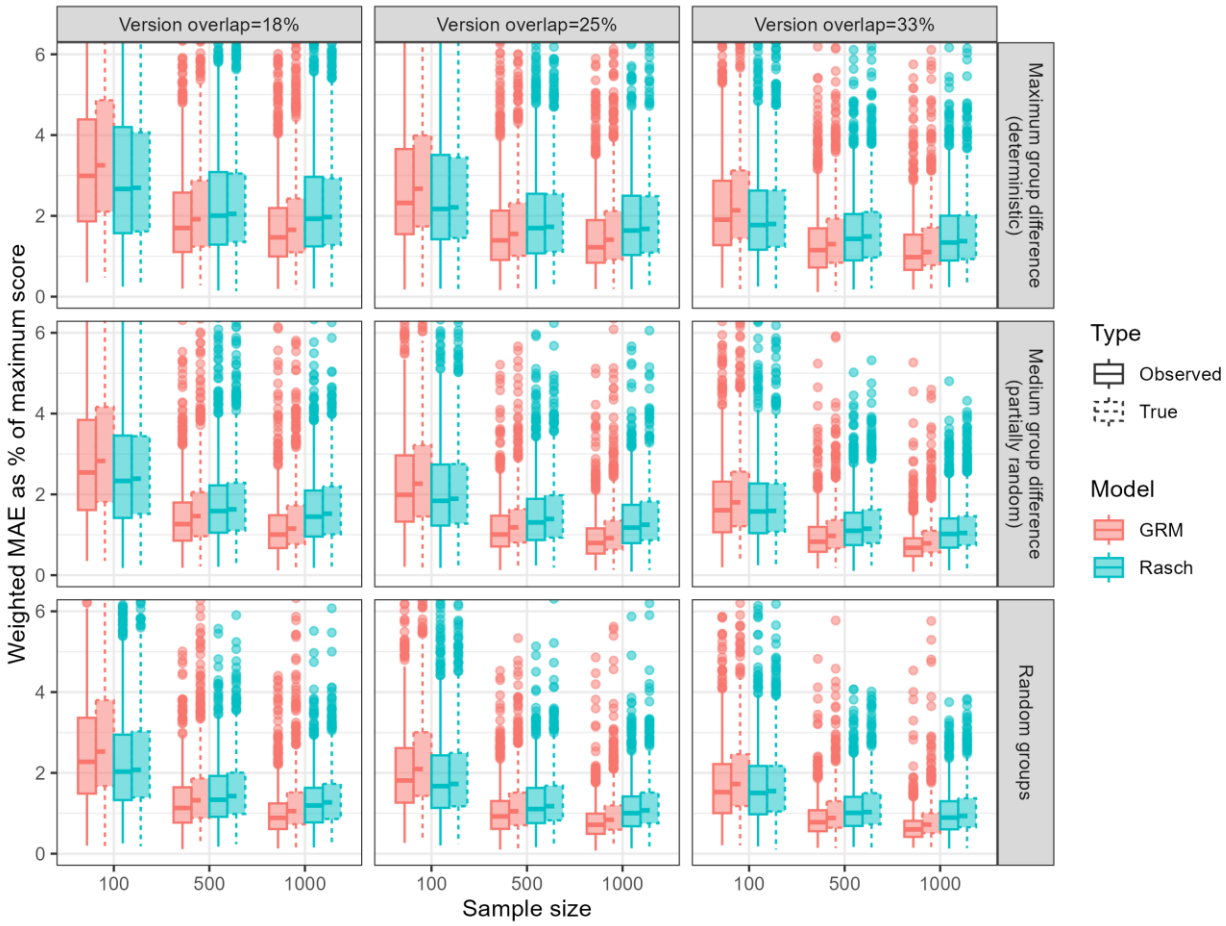

Figure 10. The distribution of weighted across a range of scenarios for true and observed score equating using either the GRM or Rasch PCM with MML estimation.

[17] Differences between fitting approaches are discussed later, but by default, GRM (and GPCM) models are fitted using MML, so we use PCM fitted via MML too, to avoid conflating separate issues.

[18] Here we consider only sample sizes of 100, 500 and 1,000, with 2,000 excluded as it produces virtually identical results to those seen in the 1,000 condition. We exclude the 40% overlap condition.

This issue is explored further in Figure 11. The figure shows the distribution (across datasets) of the difference in weighted MAE between the Rasch and GRM observed score approaches in each of the 60 scenarios. This reinforces the same findings as described above in a slightly different way. In particular, the figure shows the superior performance of Rasch equating at sample sizes below 500 (as might be expected based upon Lord, 1983). However, if we have greater overlap between test forms, the GRM approach can perform just as well as Rasch even with only 100 observations. A major caveat on this latter finding is that it relates to the Rasch model based on MML estimation only. The next subsection explores how results change once JML or CML approaches to estimating the Rasch model are applied.

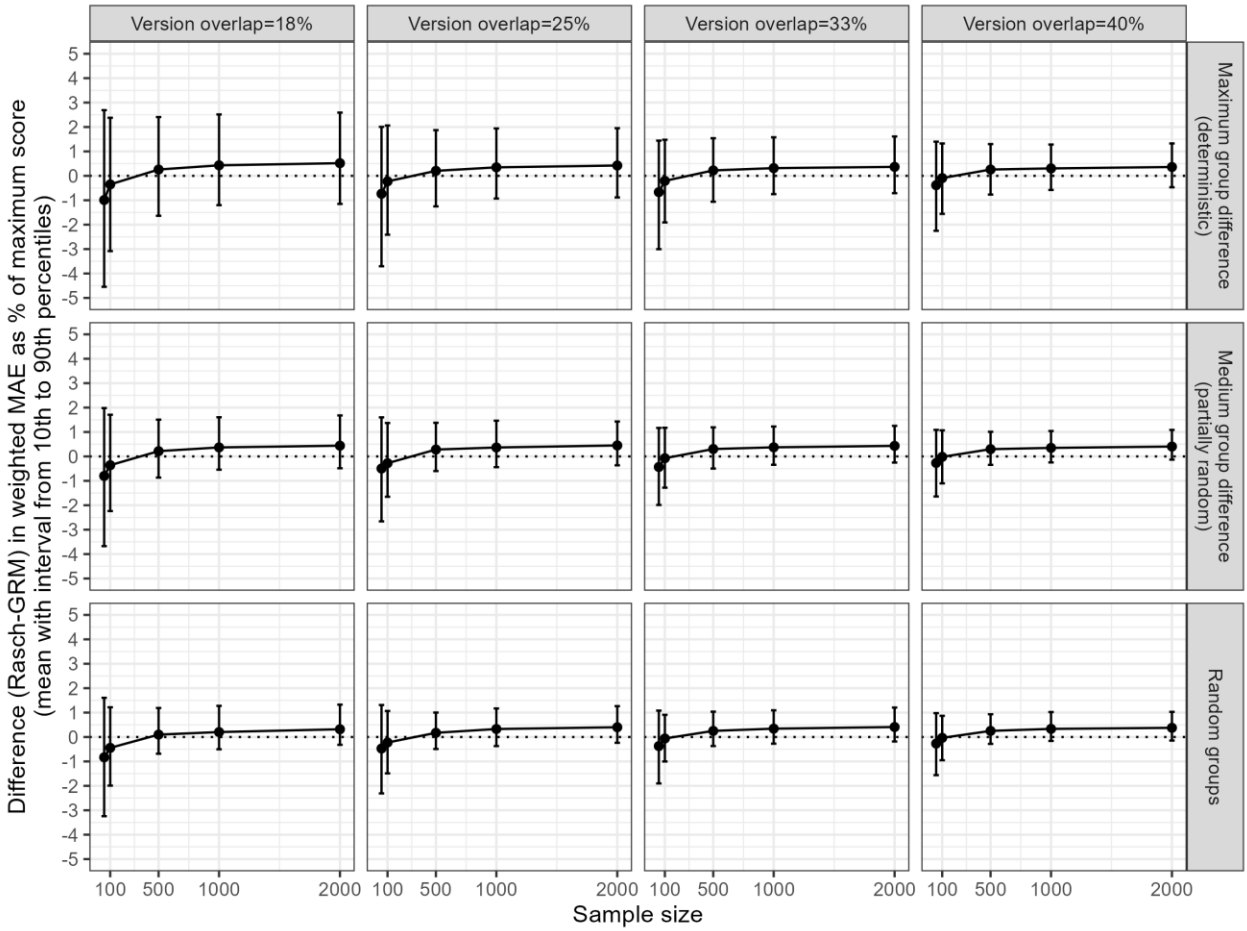

Figure 11. The distribution of differences in weighted MAE between Rasch PCM and GRM approaches to observed score equating across a range of scenarios.

JML vs CML vs MML for Rasch equating

This section compares the performance of Rasch true score equating under different model estimation approaches. True score equating is used, as observed score approaches are more difficult with JML and CML estimation. This is because observed score equating requires an estimate of the ability distribution – something the CML approach explicitly avoids and that is tricky with JML as it requires a decision about how to handle measurement error in the ability estimates of individuals. Besides, as shown in the previous subsection, for Rasch models (at least those estimated using MML) there is little difference in performance between true and observed score equating.

Figure 12 shows the distribution (across datasets) of differences in the error of equating from Rasch equating using JML and MML across all 60 scenarios. Figure 13 shows the same information, but comparing JML and CML. The fact that, in every scenario, the points in each figure are below the dotted line indicates that JML led to lower error on average than either of the other approaches. The sizes of these differences in performance are minuscule (also note the reduced scale of the y-axes). Nonetheless, the results are consistent. From this, we can conclude that *at least when completed using the default model fitting settings in equating functions in R*, use of JML leads to more accurate equating on average (albeit by a tiny margin).

Given the very small differences in performance, it is important to be clear that the results here could relate to minor technical differences (for example, convergence criteria) rather than any fundamental differences between the methods. All IRT software has a number of control parameters relating to the finer details of estimation and identifying the impact of these on equating accuracy would be challenging. The results here relate to default estimation of the Rasch model in R using the packages *mirt*, *TAM* and *dexter* for MML, JML and CML respectively. As such, they represent the most straightforward ways in which each approach could be implemented using free software at the time of writing. We add the important caveat here, then, that it may be possible to further improve outcomes for MML or CML estimation by changing parameters away from defaults. Given the relatively high complexity of this simulation study, however, we consider searching for optimal parameters for each dataset and each simulation scenario to fall outside the scope of the work. If the methods are used in practice, however, this could be a valid and fruitful step.

Noting that JML consistently performs a little better than MML gives us reason to revisit the findings in the previous subsection. In particular, if we have a sample size of 100 or less, a Rasch true score approach will be superior to an observed score approach based upon the GRM – even with a large degree of overlap with test forms. Thus, as expected, for smaller sample sizes a Rasch PCM model is preferred to any more complex approach to IRT.

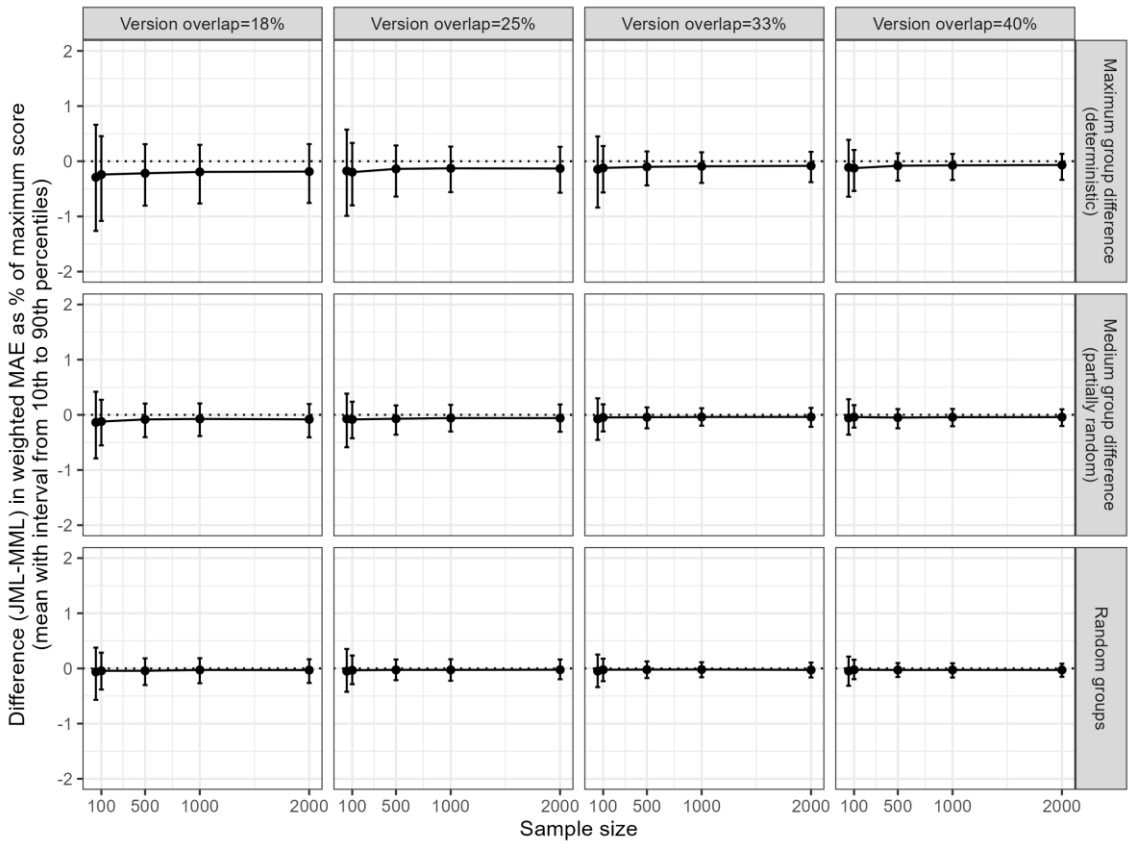

Figure 12. The distribution of differences in weighted MAE between Rasch true score equating using either JML or MML estimation across a range of scenarios.

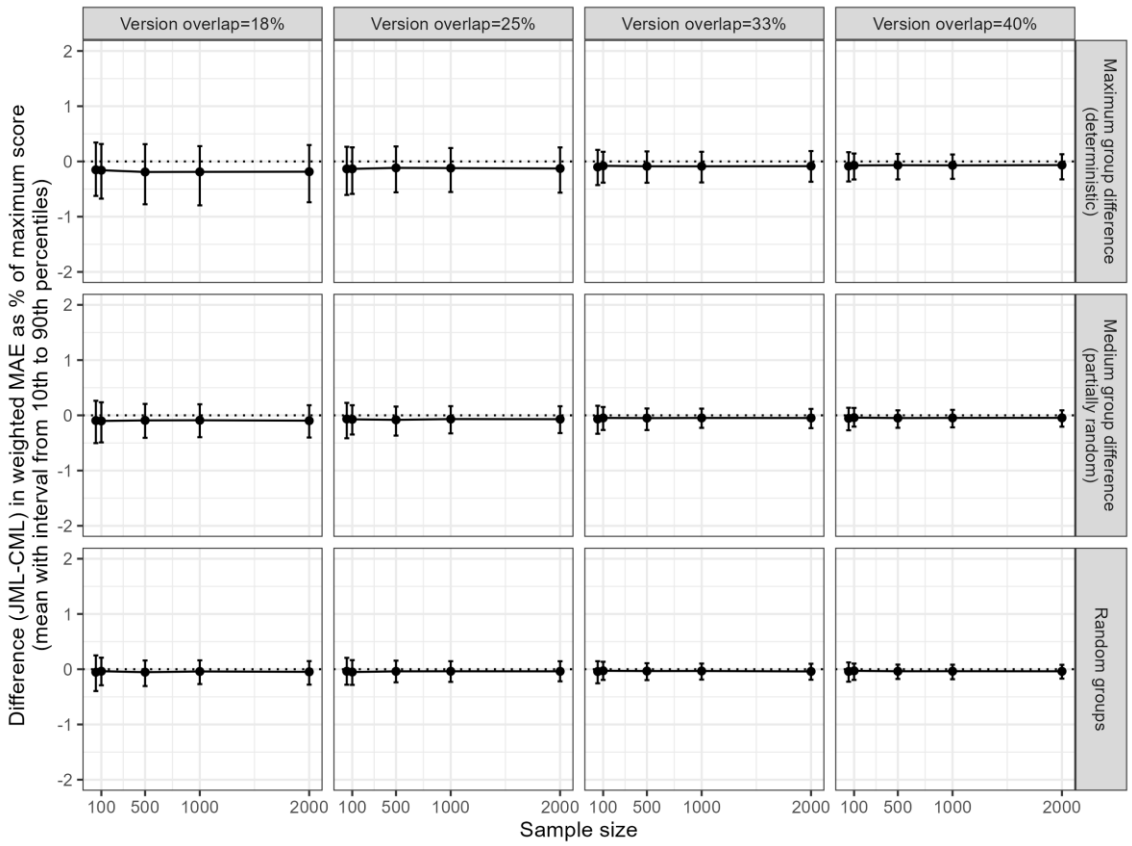

Figure 13. The distribution of differences in weighted MAE between Rasch true score equating using either JML or CML estimation across a range of scenarios.

2-parameter vs 3-parameter model

We next explore the possibility of using a more complex IRT model still. Specifically, we compare the performance of the GRM to a model where the 3PL model is used for dichotomous items and the GRM is used for all others. We focus on the performance of observed score equating because, as shown earlier, this is the most accurate approach for these more complex models. We also focus upon results from the Stocking-Lord method for each method.

The distribution of differences in error between the 3PL and pure GRM approach is shown in Figure 14 across each scenario. As can be seen, the differences are small. However, for the sample sizes below 1,000 and scenarios where we have larger differences between groups, equating based upon the 3PL model tends to have slightly larger errors (the points in the chart are above zero). That is, equating based upon a pure GRM dominates the approach including the 3PL and, in the absence of a competing justification, should always be preferred.

We note, however, that none of the assessments included in our analysis related to pure multiple-choice tests. It is possible that the performance of the 3PL would be superior in such instances. After all, such scenarios would reflect the intended application of the 3PL model, due to the increased potential for guessing in multiple choice tests, in contrast to the types of items present in the tests used in our simulation. Hence, if pure multiple-choice tests were to be equated, it may still be worth considering this approach, but it should also be compared to the use of the GRM (or, indeed, GPCM) alone.

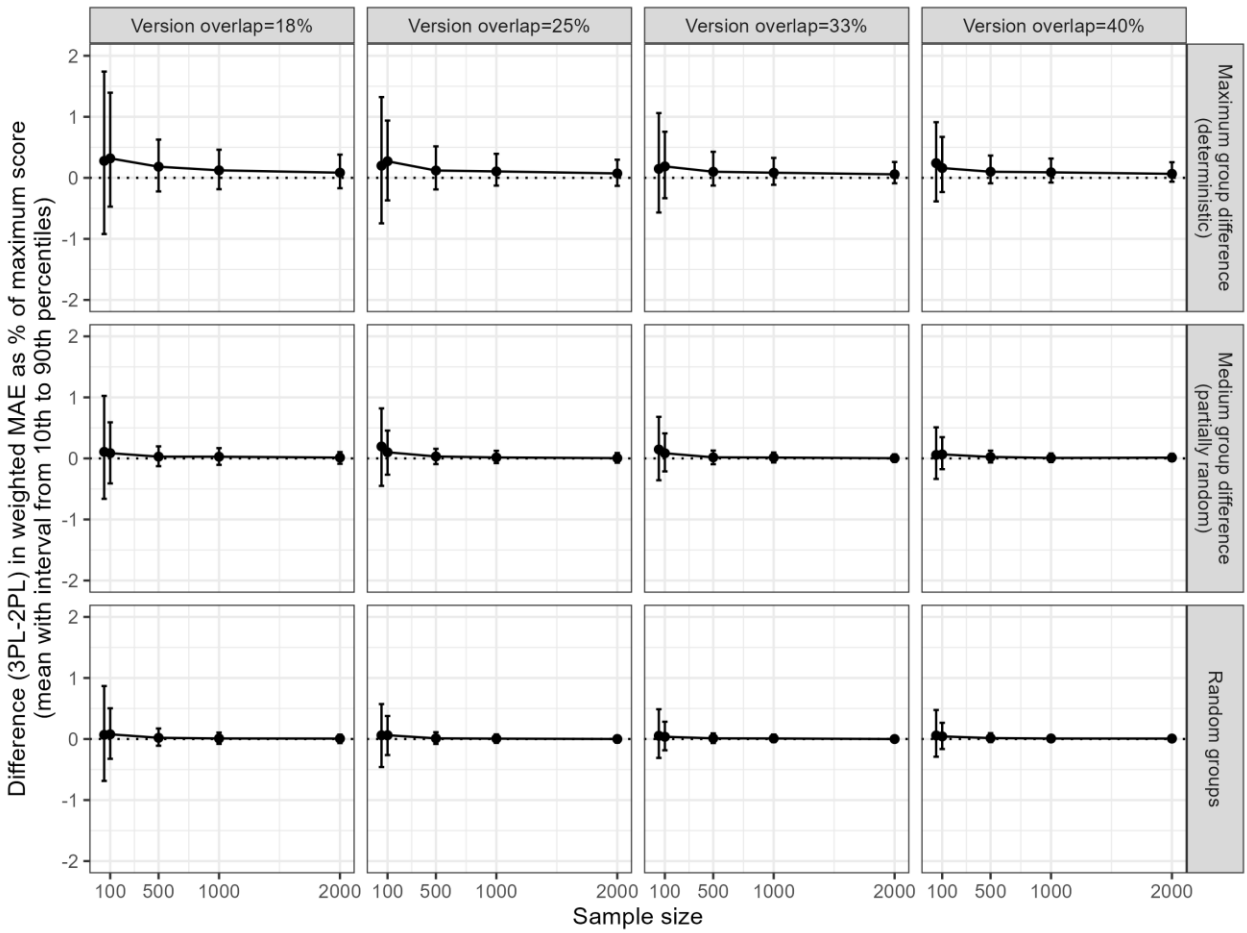

Figure 14. The distribution of differences in weighted MAE between GRM/3-parameter and GRM (2-parameter only) approaches to observed score equating across a range of scenarios.

Concurrent versus separate model estimation

Next, we consider whether it is better to put all our data into a single model and estimate all IRT item parameters once, or whether we should we estimate separate IRT models for each test form and then align the sets of item parameters using a metric transformation such as the Stocking-Lord or Haebara method. This issue has been explored previously via simulation studies by Hanson and Beguin (1999).

Figure 15 compares the distribution of the difference in error between concurrent estimation and separate estimate using the Stocking-Lord procedure for the GPCM approach. We will provide more details on the Haebara method in the next subsection. The focus in our description of results is on sample sizes larger than 500, as we have already seen that a Rasch PCM model is likely to be more effective for smaller sample sizes.

Figure 15 shows that, for sample sizes over 500, the performance of the two GRM methods was extremely similar. Having said this, except for two scenarios with the largest extent of group differences, the average level of error was always slightly lower for the concurrent approach to IRT estimation (i.e., the central points are below the dashed line). This indicates that, although the difference is tiny, it may be preferable to use a concurrent approach to IRT estimation where we have the choice. Similar findings were reported by Hanson and Beguin (1999).

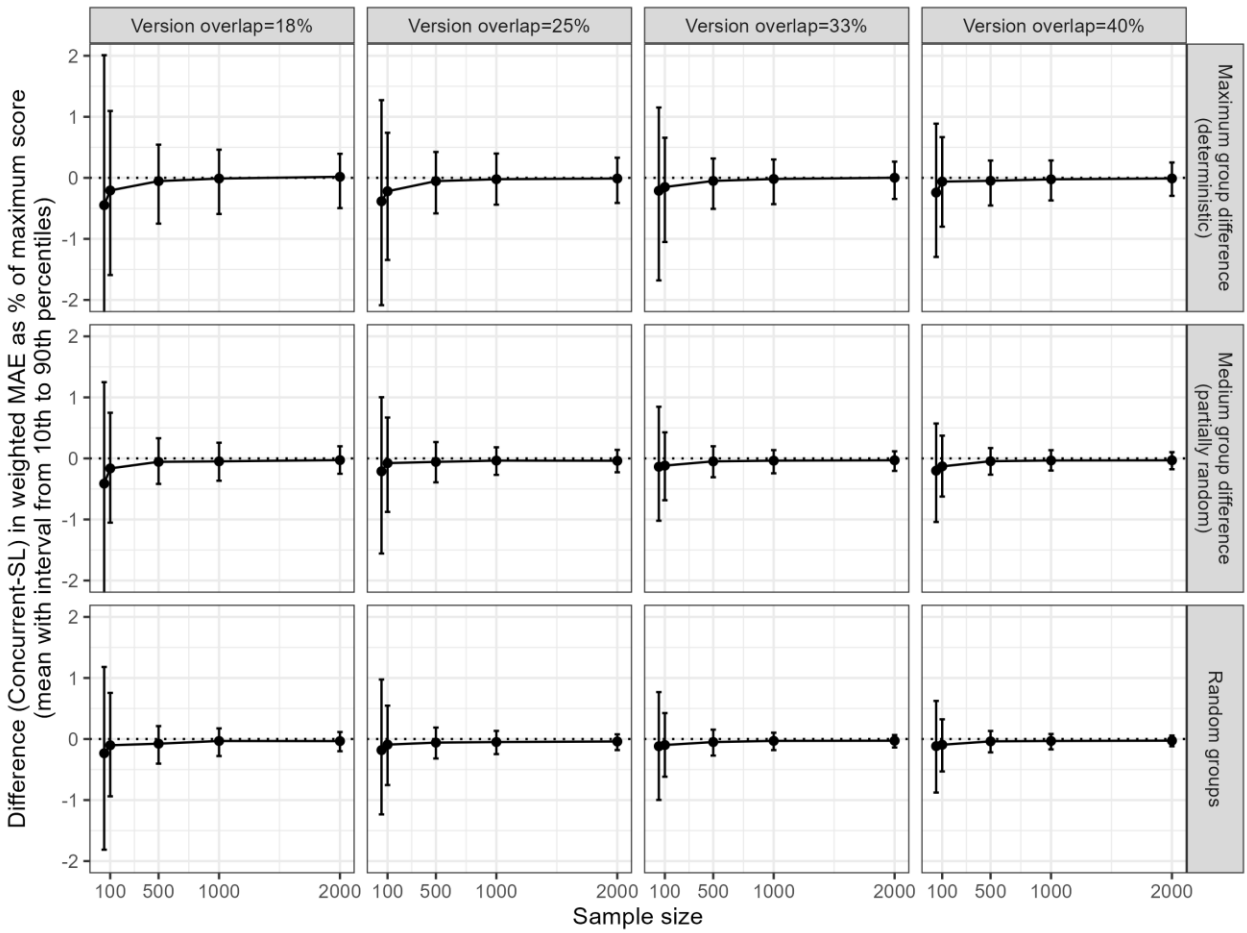

Figure 15: The distribution of differences in weighted MAE between GPCM observed score equating using either concurrent estimation or Stocking-Lord metric transformation across all scenarios.

Figure 16 shows the results for the Rasch PCM model based upon JML estimation and true score equating. The figure shows that, for all sample sizes, the concurrent approach to estimation led to slightly more accurate equating. This implies that, with the Rasch model, a concurrent approach to estimation should generally be preferred unless there are operational reasons to avoid this.

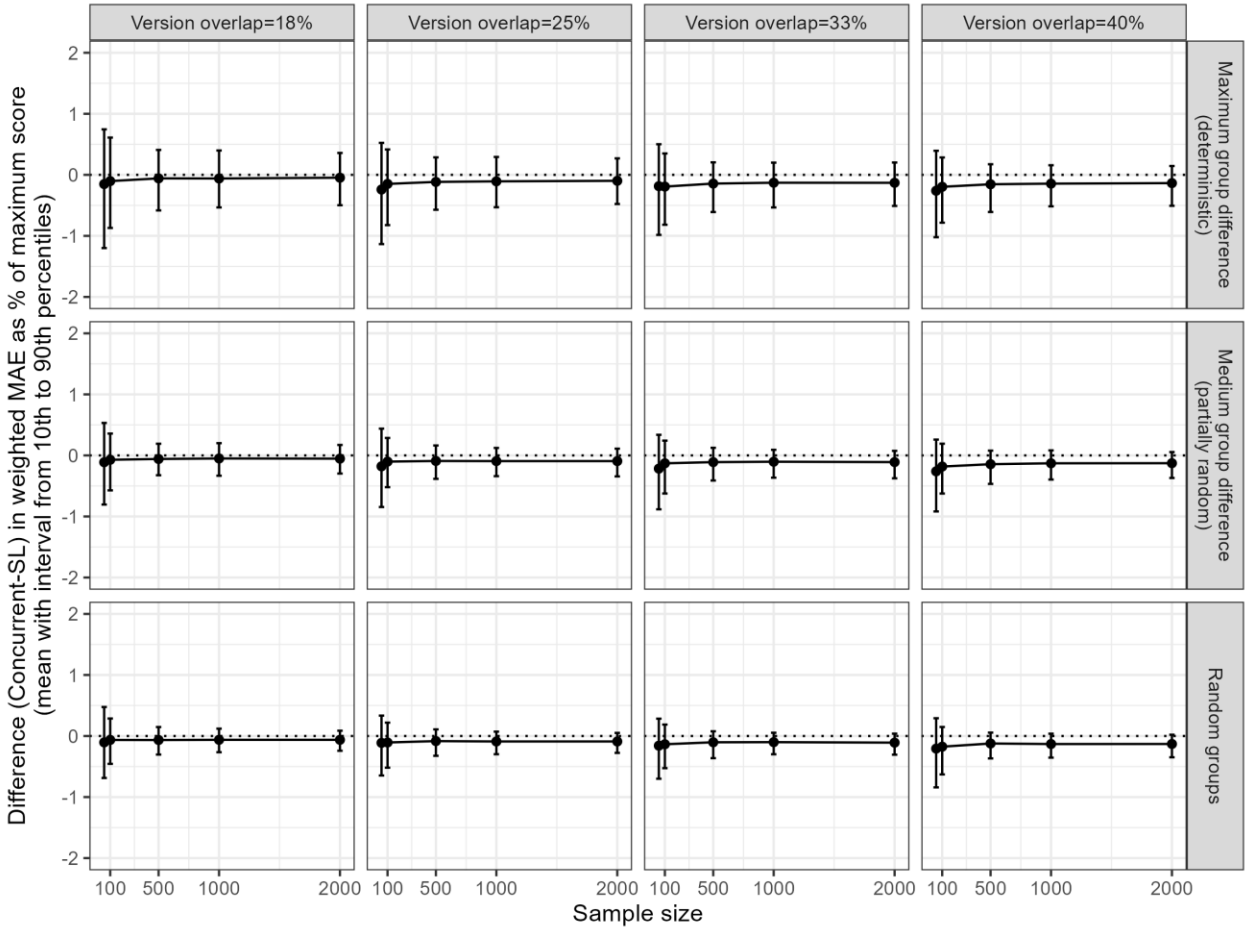

Figure 16. The distribution of differences in weighted MAE between Rasch PCM true score equating using JML estimation and either concurrent estimation or Stocking-Lord metric transformation across all scenarios.

Haebara vs Stocking-Lord

Although concurrent estimation tends to produce (slightly) better results than separate estimation, there will still be cases where separate model fitting is required, in turn necessitating a means of transforming model estimates to equate test scores. This subsection therefore examines the difference in performance of the Stocking-Lord and Haebara approaches to metric transformation. The distribution of differences in error are shown in Figure 17 for the Rasch PCM and Figure 18 for the GPCM. For the Rasch PCM, results relate to true score equating using JML estimation, as this was found to be the most accurate on average. In contrast, the results for the GPCM relate to observed score equating (and MML estimation).

As with many results in this section, the differences in the performance of the two methods are tiny. This reflects the findings of Hanson and Beguin (1999). However, in every scenario, the Stocking-Lord approach leads to lower equating error on average. Overall, then, although the differences are small, the results suggest that the Stocking-Lord rather than the Haebara method should be preferred.

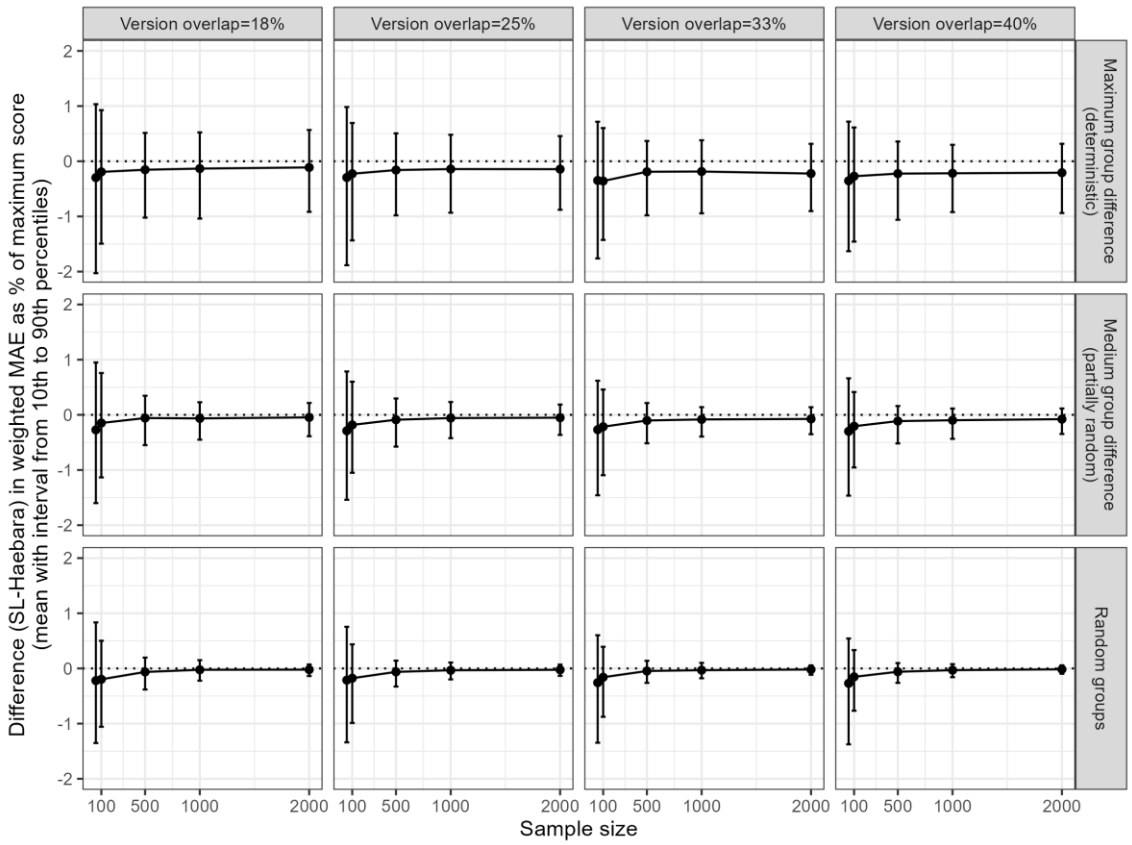

Figure 17: The distribution of differences in weighted MAE between Rasch PCM true score equating using either the Stocking-Lord or Haebara metric transformation across a range of scenarios.

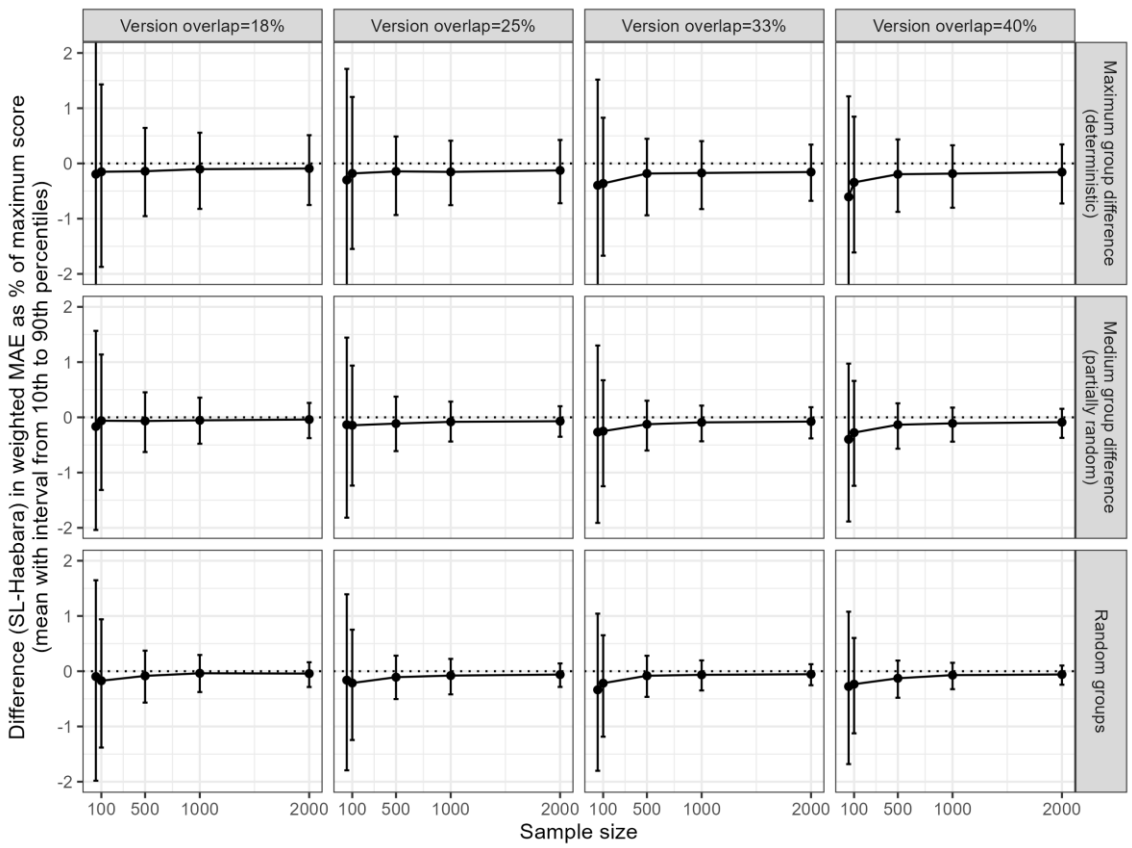

Figure 18. The distribution of differences in weighted MAE between GPCM observed score equating using either the Stocking-Lord or Haebara metric transformation across all scenarios.

Using IRT equating when the model doesn't fit

Our final question here relates to accuracy of IRT equating when the data suggests that the model does not fit the data. The fact that an IRT model does not provide a perfect fit to a given dataset is entirely to be expected. As noted by Rasch (1960) "no models are [true] – not even the Newtonian laws". He went on to say, "Models should not be true, but it is important that they are applicable, and whether they are applicable for any given purpose must of course be investigated" (pages 37-38)[19]. The purpose of this subsection is to consider exactly how wrong an IRT model can be before it ceases to be applicable. That is, since all analyses are based upon real datasets, we can get a sense of how well IRT equating work in practical situations where model fit is less than ideal.

For the purposes of this analysis, we quantify model fit in two ways. First, we use the Tucker-Lewis index (TLI). This metric, which is drawn from structural equation modelling, is based on the reduction in the overall chi-square-like[20] model error per degree of freedom when compared to a null model that assumes complete independence of all item scores. This reduction is compared to the reduction that would be achieved by an ideal model (which would have chi-square-like error of 1 per degree of freedom). According to Cai et al (2023) the TLI can provide a useful metric of overall model fit for IRT models. Ideally the TLI should have a value close to 1 and according to Shi et al (2019) values greater than 0.95 indicate a good model fit[21].

In addition to the TLI, we also explore model fit based upon the standardised root mean square residual (SRMSR) as recommended by Maydeu-Olivares (2013). This metric is simply based on the root mean square differences between the observed pairwise correlations between item scores and the correlations that would be expected if the IRT model was correct. Maydeu-Olivares suggests that the SRMSR should be below 0.05 for well-fitting IRT models.

For each of the 1,000 datasets in the analysis, IRT fit was calculated based upon fitting a model to the full-length test (i.e., before it had been split into 2 forms) for a random sample of 2,000 candidates[22]. IRT fit was calculated separately for the GRM and Rasch PCM models. Fit calculations completed without error for 898 of the 1,000 datasets and so analysis in this section is restricted only to these.

For the purposes of this section, we look at the distribution of the differences in error between IRT and chained equipercentile equating with kernel smoothing. This version of chained equating is used as a reasonable point of comparison as results presented above show it performs fairly well across a range of sample sizes. For the GRM we focus upon observed score equating and for the Rasch PCM we focus upon true score equating based upon JML parameter estimates. For both IRT approaches we report results based upon concurrent estimation.

Figure 19 shows the distribution of differences in error between chained equipercentile and IRT equating via the GRM in each scenario. Different coloured lines and differently shaped points

---

[19] Reproduced from https://en.wikipedia.org/wiki/All_models_are_wrong#Historical_antecedents.

[20] Technically, for IRT models, the M2 statistic (see Cai, 2023).

[21] Cai (2023) states that a reasonable cutoff should be above 0.97.

[22] A random sample was used at this point in analysis to speed up calculations.

are used to distinguish situations where the TLI exceeds the recommended level of 0.95 (the red line), where it fails to meet this cut-off but is at least above 0.90 (the green line), and where the TLI is below 0.90. Note that 70% of datasets had a TLI in excess of 0.95, 20% had a TLI between 0.90 and 0.95 and the remaining 10% had a TLI below 0.90. We would expect the performance of GRM equating to be best when we have the best IRT fit (i.e., the red lines should be the lowest). The expected pattern is clearly seen in the scenarios with the largest group differences (i.e., the top row), but is much less evident in the scenarios with smaller differences between groups. Indeed, where we have non-random ability differences between groups, and the sample size is at least 500, GRM equating is usually more accurate than chained equating even in cases where the TLI was below 0.90. This indicates that a lack of model fit generally has only a moderate impact on equating accuracy, and should not alter our overall conclusions about the suitability of IRT equating.

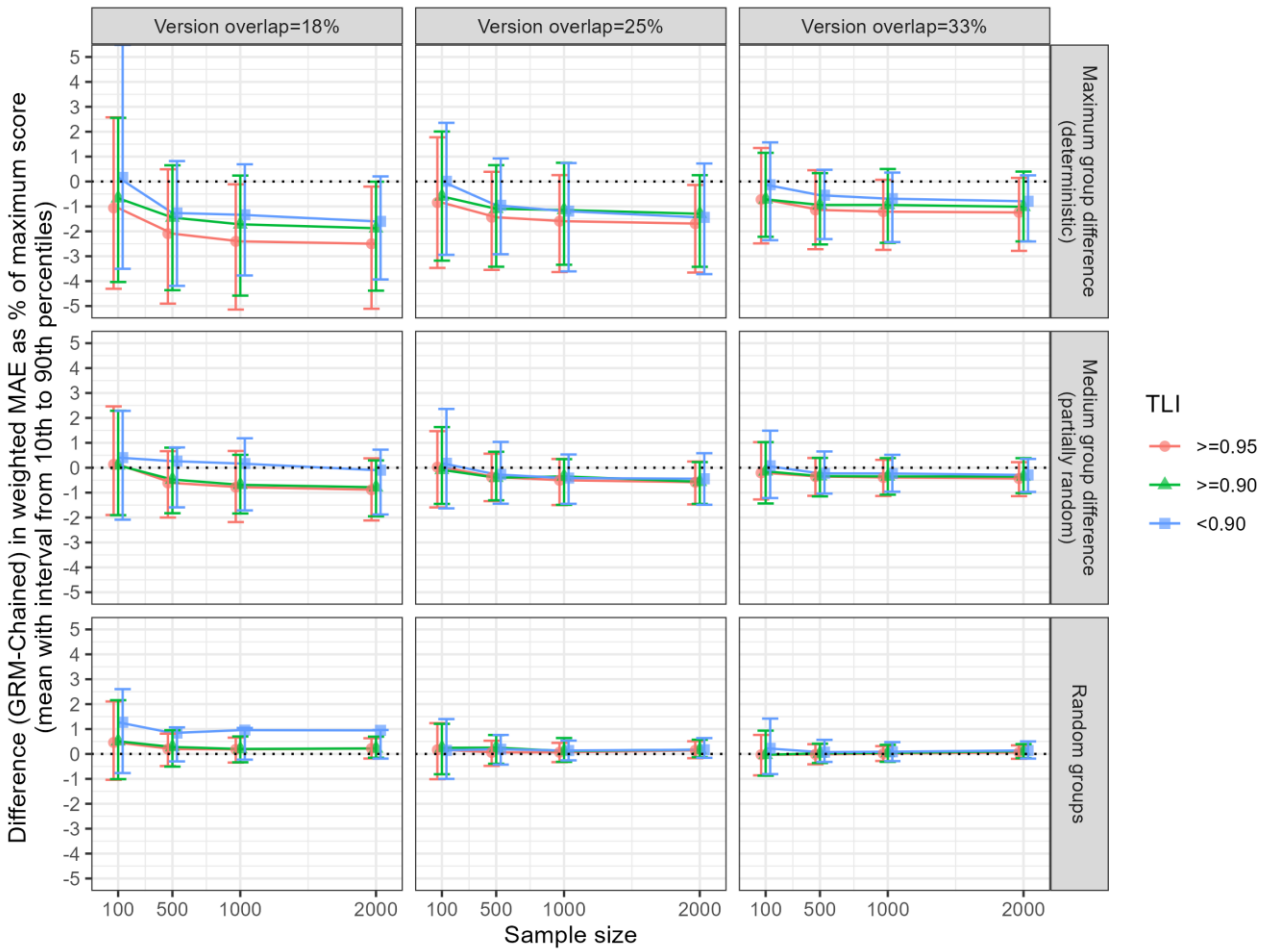

Figure 19. The distribution of differences in weighted MAE between GRM observed score equating and chained equipercentile equating with kernel smoothing across a range of scenarios split by IRT model fit to the original full test dataset in terms of the TLI.

Figure 20 shows the same analysis but this time the datasets are split by whether the SRMSR was below the recommended level of 0.05 (the red lines) or higher than this (the blue lines). For the GRM no datasets had an SRMSR value above 0.10. Just 14% of datasets had an SRMSR value above 0.05 indicating that the majority showed a good fit to the GRM.

In every scenario, the GRM model outperforms chained equipercentile equating most clearly where we have a good model fit as measured by SRMSR. This indicates that the SRMSR is a better indicator of model fit than the TLI if our goal is to ascertain whether model fit is sufficiently good for the purposes of equating. Nonetheless, Figure 20 also shows that, even where model fit was worse than ideal, if there were non-random ability differences between groups and a sample size of at least 500, on average, IRT equating was still more accurate than a chained equipercentile approach.

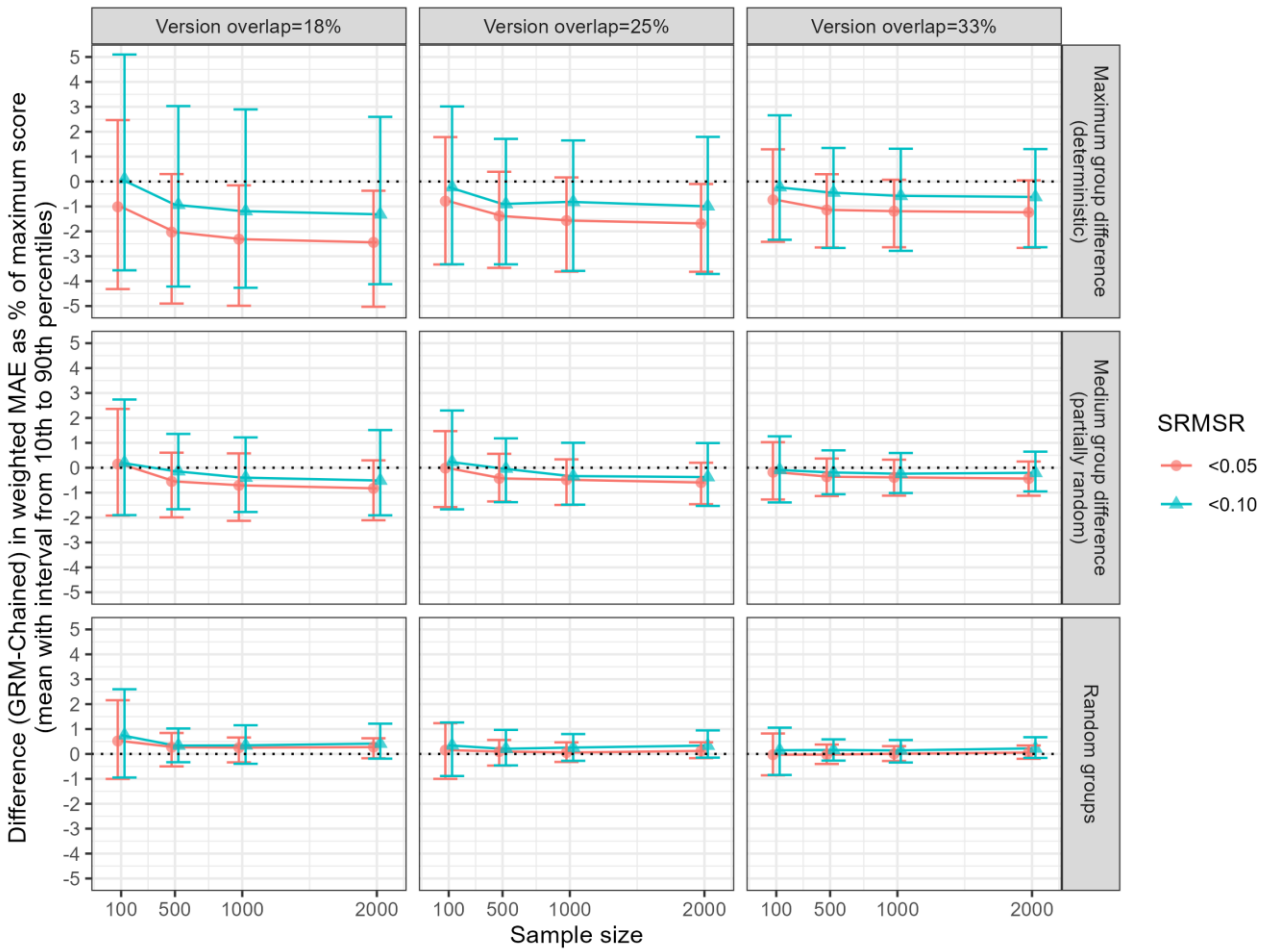

Figure 20. The distribution of differences in weighted MAE between GRM observed score equating and chained equipercentile equating with kernel smoothing across a range of scenarios split by IRT model fit to the original full test dataset in terms of the SRMSR.

Figure 21 repeats the TLI analysis for the Rasch JML approach. The accuracy of IRT equating is compared to that of chained equipercentile equating, with results split by the TLI of the Rasch PCM model in different datasets. For the Rasch PCM, 45% of datasets had a TLI larger than 0.95, 36% had a TLI between 0.90 and 0.95, and 20% had a TLI below 0.90. The results were similar to those seen for the GRM. The performance of IRT equating (relative to chained equipercentile) tended to be slightly better where we have the higher TLI values, with this particularly evident in scenarios with the largest group differences. However, for non-random group ability differences, Rasch equating tended to be superior to chained equipercentile equating even if model fit was less than ideal.

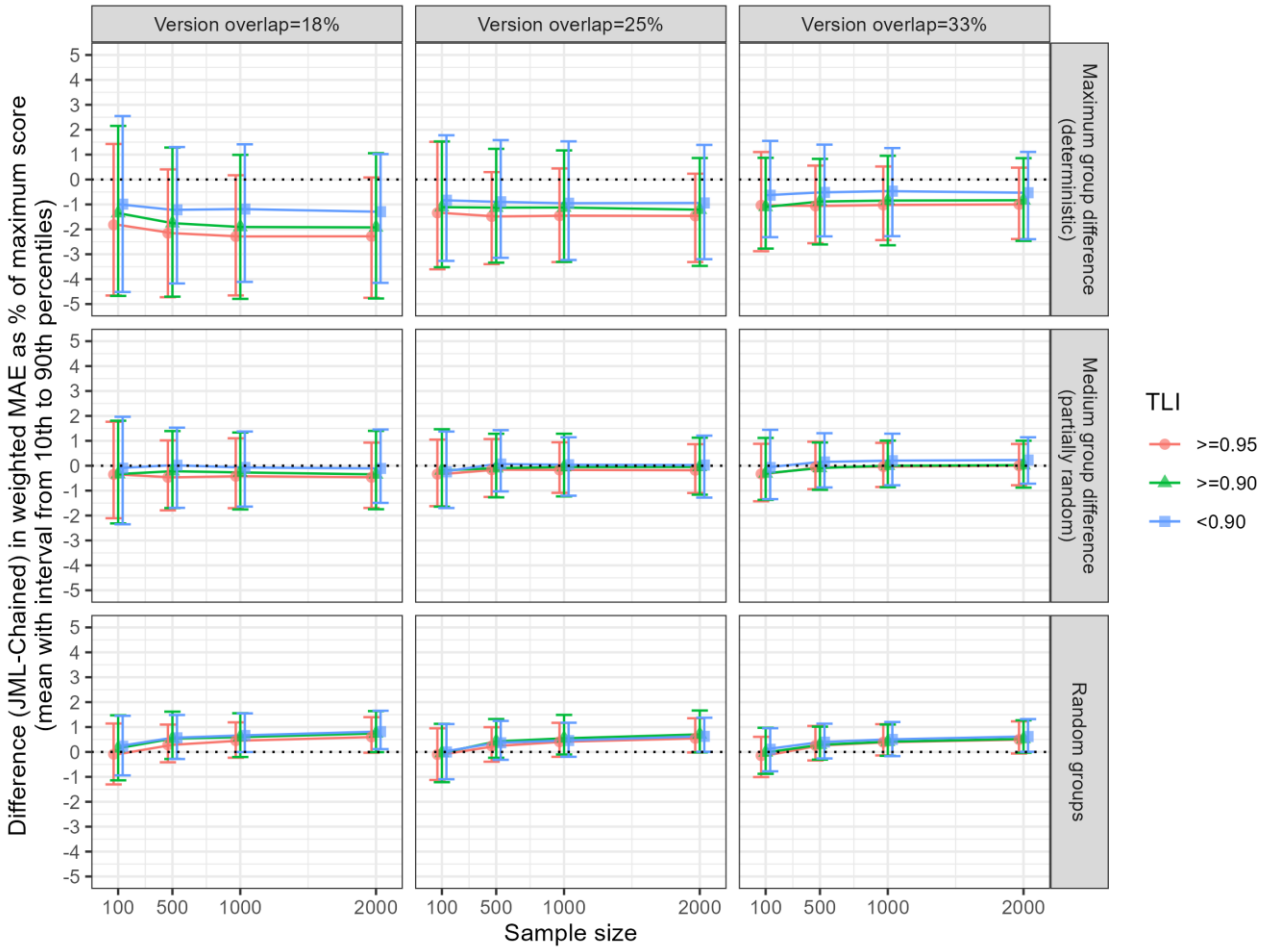

Figure 21. The distribution of differences in weighted MAE between Rasch PCM true score equating and chained equipercentile equating with kernel smoothing across a range of scenarios split by IRT model fit to the original full test dataset in terms of the TLI.

Finally, Figure 22 repeats the SRMSR analysis, but for the Rasch PCM (JML) approach. For the Rasch PCM, the vast majority of datasets (91%) had SRMSR values between 0.05 and 0.10. Only 3% had SRMSR values below 0.05 meeting the criterion set by Maydeu-Olivares (2013) for a good fit. The remaining 6% had SRMSR values in excess of 0.10. There was a slight (but by no means universal) tendency for the performance of Rasch equating relative to chained equating to be best when the model displayed the best fit. However, in scenarios with non-random differences between groups, even when the SRMSR was greater than 0.10, Rasch true score equating tended to perform close to or slightly better than a chained equipercentile approach.

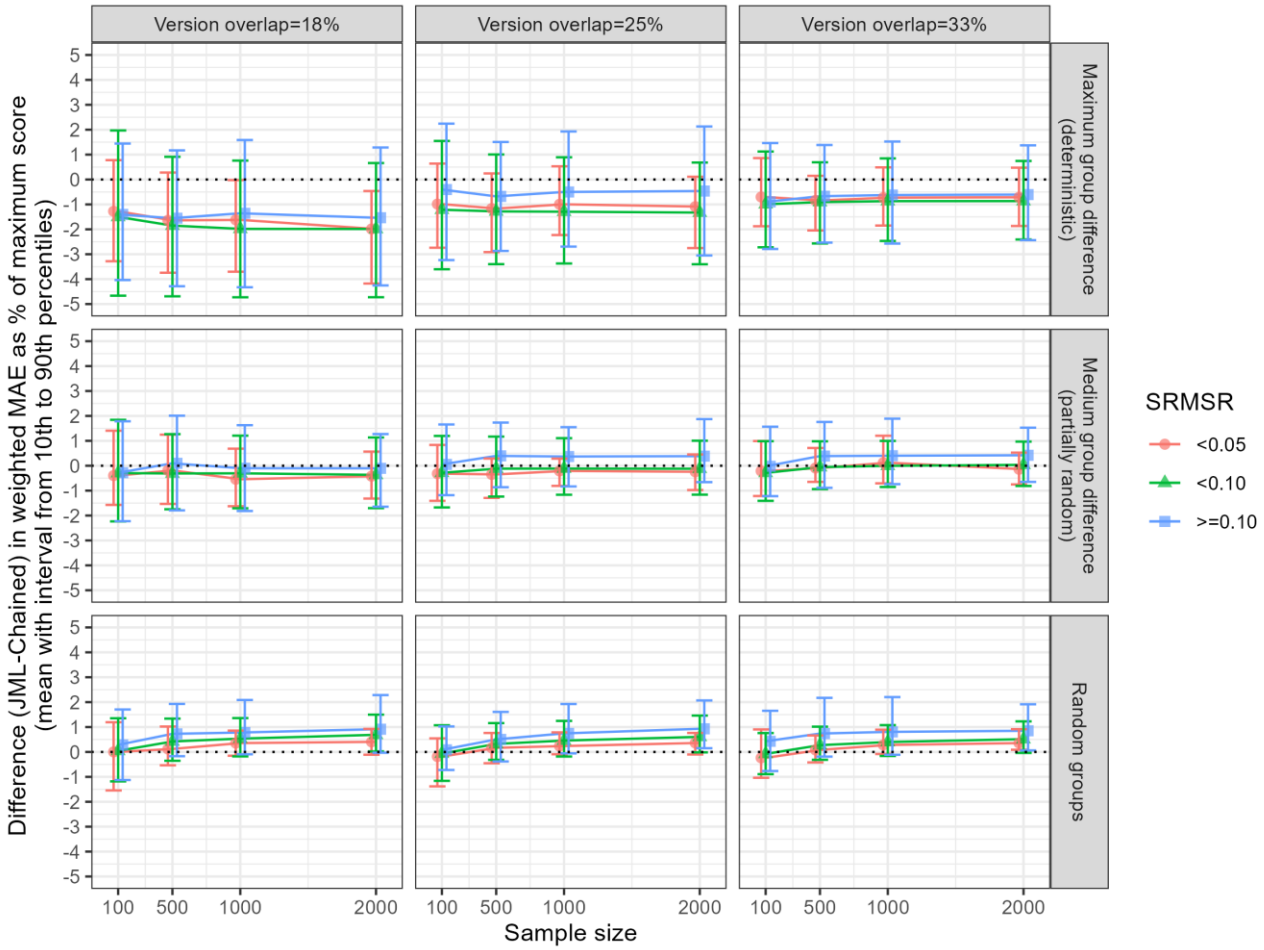

Figure 22. The distribution of differences in weighted MAE between Rasch PCM true score equating and chained equipercentile equating with kernel smoothing across a range of scenarios split by IRT model fit to the original full test dataset in terms of the SRMSR.

Therefore, in summary, this subsection has shown that IRT equating can provide an accurate approach even when the model does not meet some of the criteria for model fit that are recommended in the IRT literature. As shown earlier, there will clearly be conditions where use of IRT is *not* recommended, most notably situations with very small sample sizes. However, once an appropriate sample size is reached, it seems that IRT equating is more robust than might be expected, and thus can still outperform classical methods even with worse-than-ideal model fit.

Secondary questions about classical equating

In the previous section we focused on issues around IRT equating, so here we (primarily) dig deeper into comparisons between different classical equating techniques.

Rasch JML vs the best performing classical method in each scenario

Before considering the various classical equating techniques, we first consider the performance of an IRT approach, namely Rasch PCM true score equating with JML estimation, relative to the best classical approach in each scenario. In each scenario, we identified the classical technique with the lowest average weighted MAE across datasets (see Table 6**Error! Reference source not found.**), then compared its performance to that of the PCM (JML Concurrent True) approach. Figure 23 shows the distribution of differences in weighted MAE between Rasch JML and the best performing classical method. Values above zero indicate that a classical method outperforms the PCM whereas values below zero represent the PCM outperforming every possible classical method.

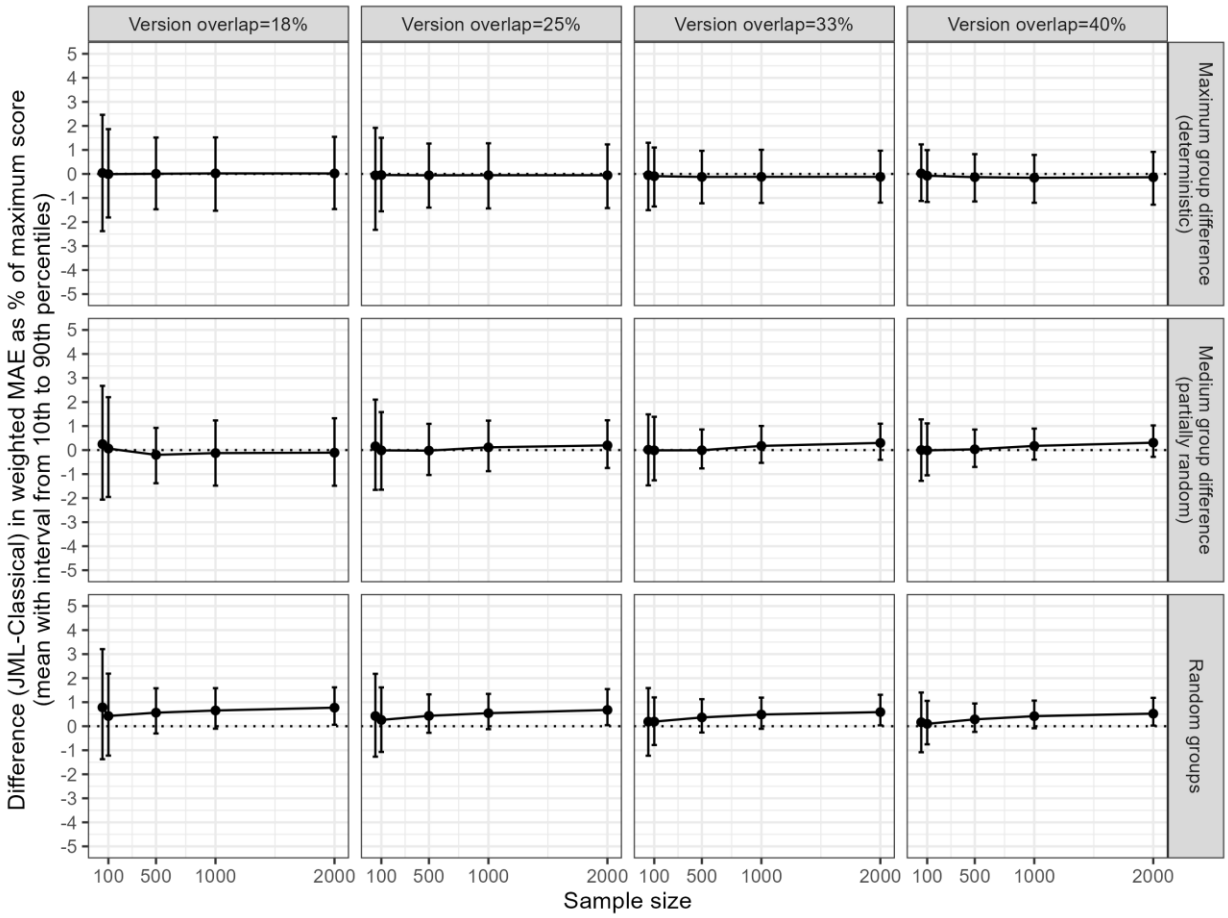

Figure 23. The distribution of differences in weighted MAE between Rasch PCM true score equating and the best performing classical method across all scenarios.

The error of Rasch PCM (JML) equating is never far from the error of the best performing classical method. This indicates that the Rasch approach is not a bad default method to rely on. That said, a well-chosen classical equating technique will never be substantially outperformed by the Rasch approach. For example, Table 5 showed earlier that for a sample size of 100, and medium or large group differences, the Rasch approach was the best performing method, but from Figure 23 we can see that, in these scenarios, the difference with the best performing classical method is barely visible.

Levine vs Tucker vs Chained (linear)

Next, we compare three common linear methods. To begin with, Figure 24 shows the distributions of differences in error between chained and Tucker linear equating. The results show that, unless group differences are random, chained equating is clearly superior (i.e., the points are all comfortably below the central line of equality). Even in the bottom row of the figure, which shows results for random groups, the improvement in performance from using Tucker rather than chained equating is extremely small. With this in mind, chained linear equating can almost always be used in preference to a Tucker approach. Although charts are not included here, similar arguments could be made to prefer chained linear equating to the Braun-Holland and PSE variants of linear equating.

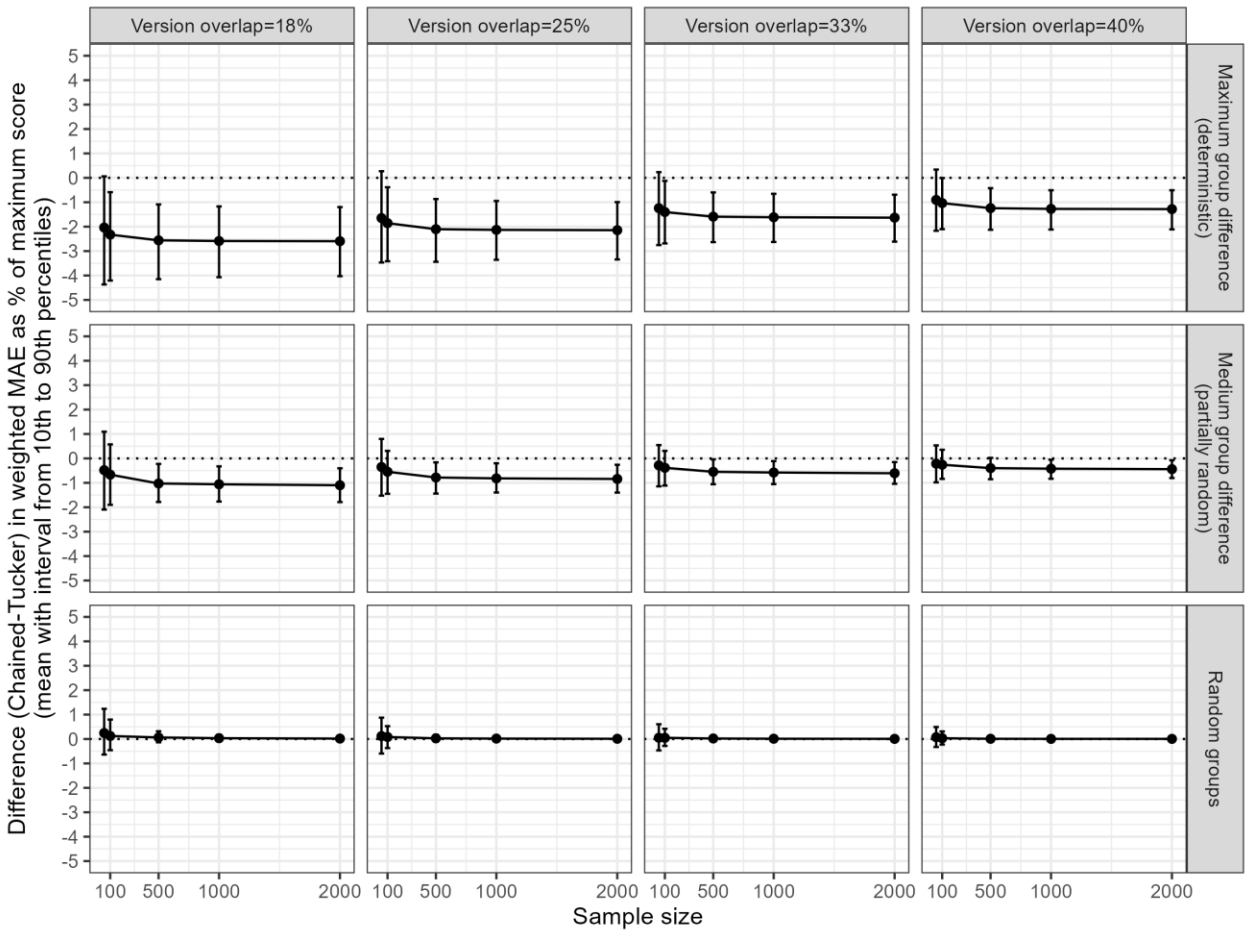

Figure 24. The distribution of differences in weighted MAE between chained linear and Tucker linear equating across a range of scenarios.

Amongst linear approaches, this leaves a choice between the chained and Levine variants. Figure 25 shows the distribution of differences in error between these two methods in each scenario. Points above the central horizontal line indicate that Levine equating is superior and points below it indicate that chained equating is superior. The superior method clearly varies across different scenarios. For the largest differences between groups, the Levine approach is superior except for sample sizes below 100. For scenarios with medium group differences, chained is superior when the sample size is less than 500. For purely random group differences, chained equating is always superior.

Note that both chained and Levine equating have equipercentile equivalents. In place of chained linear equating, we could use chained equipercentile equating (with kernel smoothing) and in in place of the linear Levine method we could use the Modified Hybrid approach (which is built on a combination of Levine linear equating and smoothed equipercentile equating). Although, for brevity, this comparison is not shown here, the comparison between these two equipercentile approaches yields very similar results to the comparison of chained and Levine linear methods.

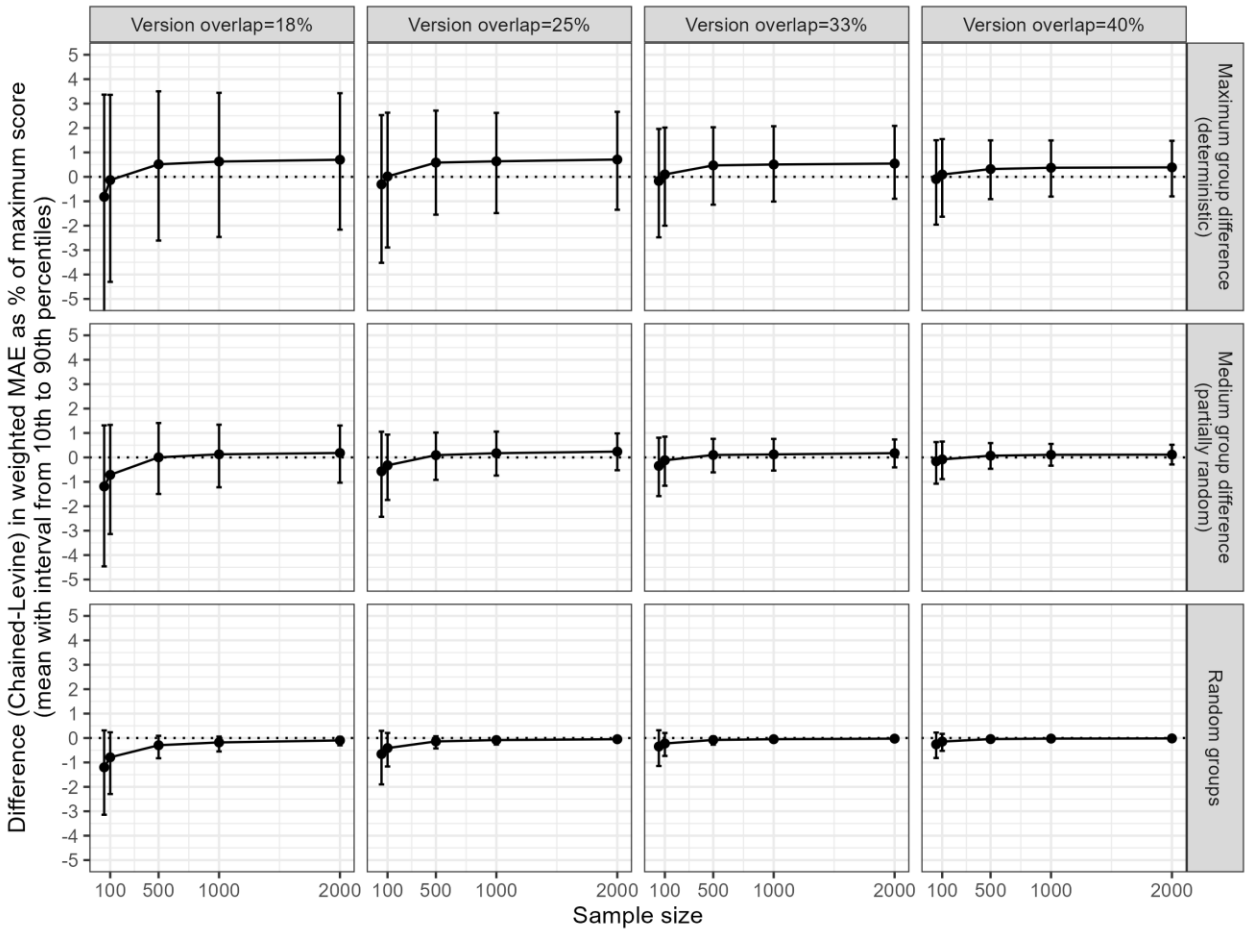

Figure 25. The distribution of differences in weighted MAE between chained linear and Levine observed score linear equating across a range of scenarios.

Symmetric circle-arc Levine vs linear Levine

Whilst considering Levine-type methods, it is worth comparing the linear Levine method to a circle-arc approach based on the Levine approach to estimating group means on each test form within the joint population of candidates. Exploring this comparison in detail is of interest given the surprisingly good performance of circle-arc equating in the original analysis run.

Figure 26 shows the distribution of differences in error between symmetric circle-arc and linear Levine approaches within each scenario. Unsurprisingly, given that it is intentionally designed as a method for use in small samples, the circle-arc approach is superior for small sample sizes regardless of the size of the differences between groups or the length of the anchor test. What is more surprising is that, if we have large differences between groups, the circle-arc approach seems superior regardless of sample size.

The reasons for this are not clear but may relate to the fact that the scenarios with large differences between groups are the most challenging scenarios explored within our analysis. As such, given the difficulties of estimating equating relationships, it may be that the very simple nature of the circle-arc approach, where we only need to estimate what the mean scores of the two test forms would be in a particular population, is beneficial. These results illustrate why we have included several methods with this same property within our simulations; we will return to comparisons of these approaches shortly.

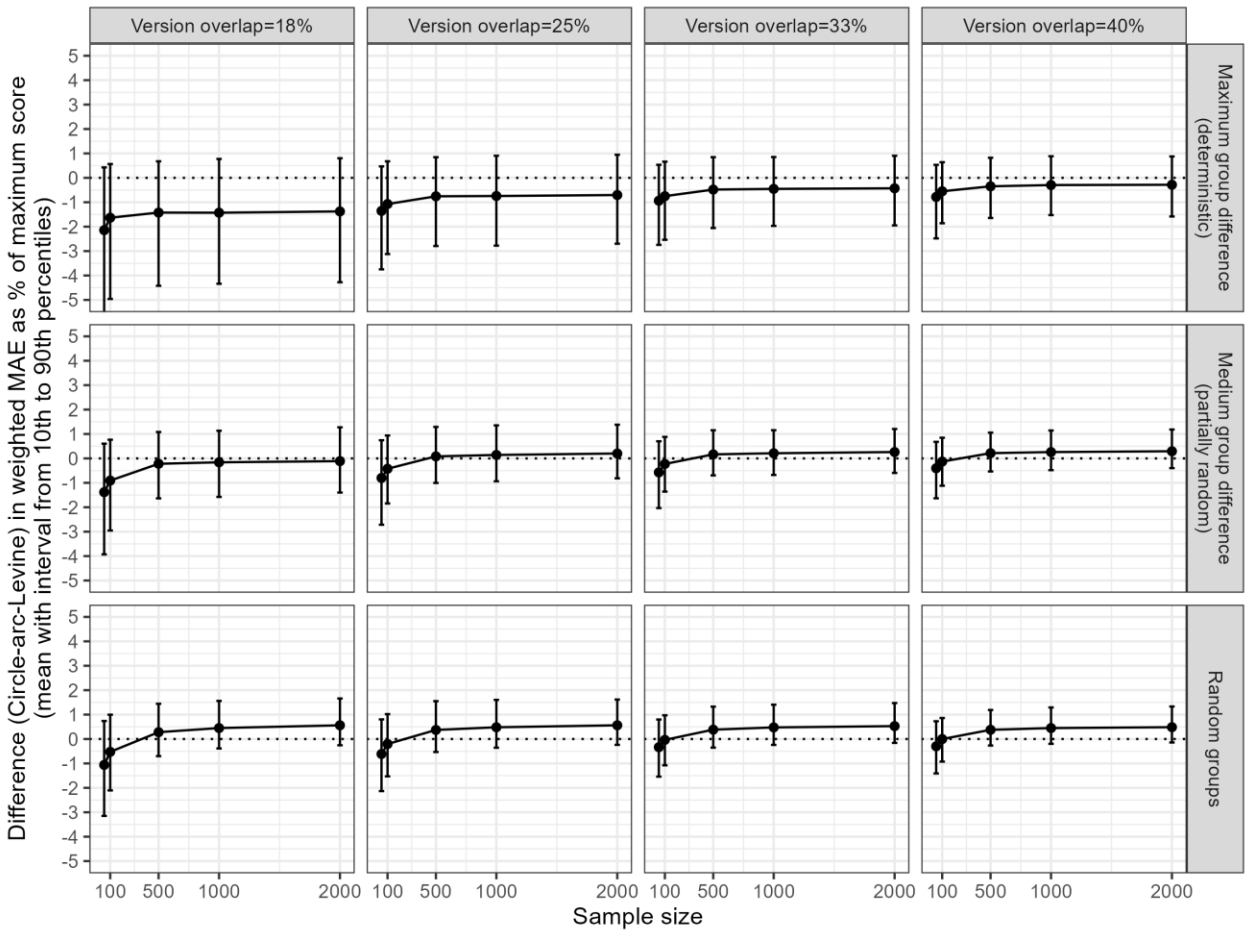

Figure 26. The distribution of differences in weighted MAE between symmetric circle-arc Levine equating and Levine linear equating across a range of scenarios.

The impact of transformations before (chained) linear equating

Reviewing earlier work in this section we can ascertain that, of the linear equating methods, chained linear equating is the only one truly worthy of consideration: Tucker equating never performs substantially better than chained equating and can perform substantially worse; Levine linear equating *can* perform substantially better than chained linear equating but *only in circumstances where we should use other methods* (such as circle-arc or odds-transform equating, not to mention IRT approaches).

With this in mind we explore the performance of chained linear equating depending upon whether we apply a non-linear transformation to scores before applying the technique (as described in the Methods section). Figure 27 shows the distributions of differences in error between a pure linear approach and one including a transformation. In every case, the points are above the central horizontal line, indicating (as noted earlier) that applying a transformation before applying equating dominates the purely linear approach. This final piece of evidence therefore suggests that there are no obvious circumstances where (untransformed) linear equating should be used.

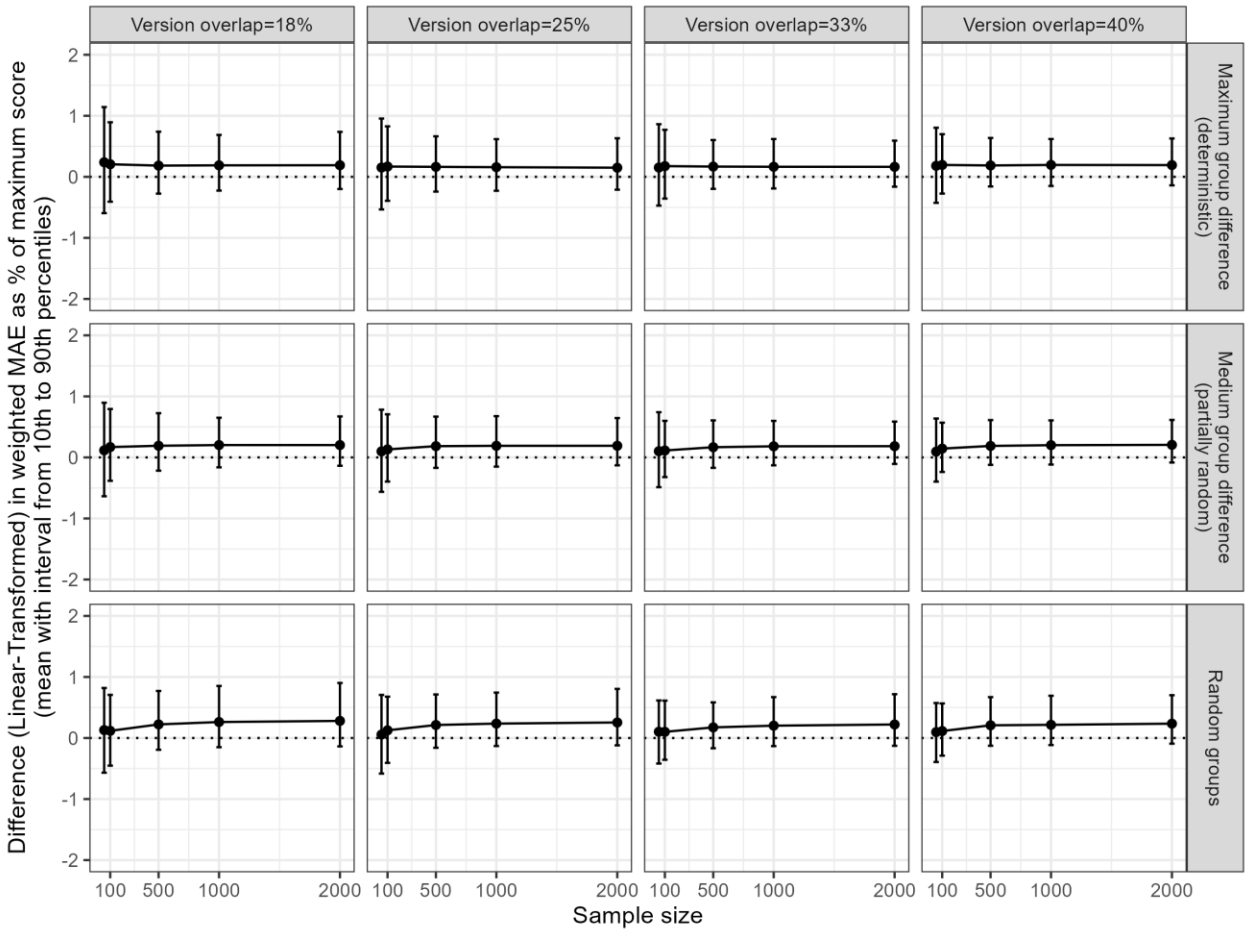

Figure 27. The distribution of differences in weighted MAE between standard chained linear equating and chained linear equating with an initial score transformation a across a range of scenarios.

Smoothed vs unsmoothed chained equipercentile

Ignoring small-sample and linear techniques, perhaps the most common classical approach to equating is the equipercentile approach. Figure 28 shows the distribution of the differences in error between unsmoothed and kernel-smoothed chained equipercentile equating. As can be seen, in every scenario, the points are above the central horizontal line, indicating that the approach including smoothing dominates the unsmoothed approach. This is particularly evident at small sample sizes. For this reason, it is advisable for equipercentile approaches to *always* include some form of smoothing. Note that, although differences are much smaller, equipercentile equating based on kernel smoothing also dominates the various alternative approaches to smoothing (i.e., loglinear and bump smoothing).

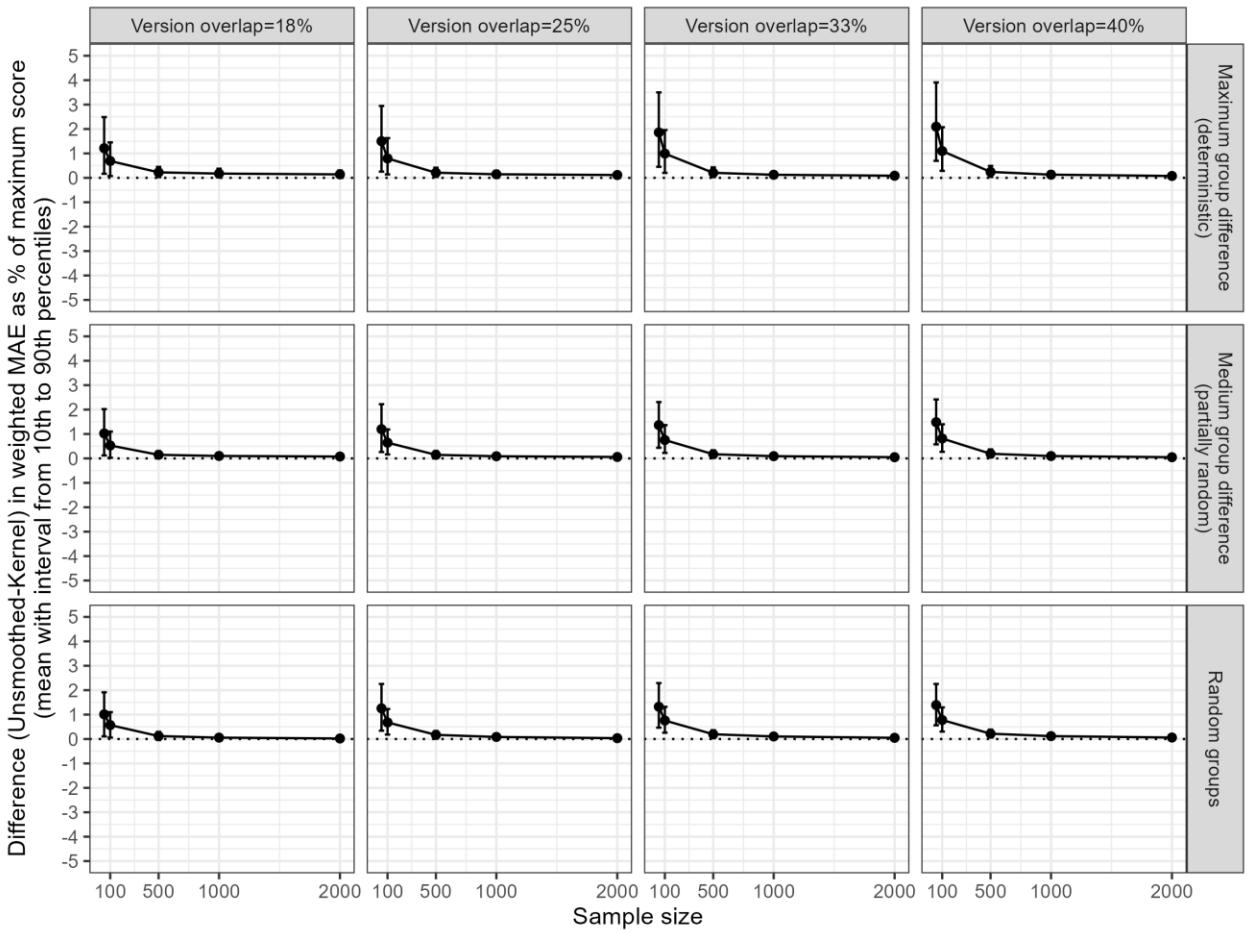

Figure 28. The distribution of differences in weighted MAE between unsmoothed and kernel smoothed chained equipercentile equating across a range of scenarios.

Smoothed equipercentile vs transformed linear (chained equating)

Earlier, we established that chained linear equating is the only form of linear equating worth considering. Furthermore, we have established that if we want to apply chained linear equating, it is always better to apply a score transformation before applying the technique. We have also established that if we wish to use chained equipercentile equating, an approach incorporating kernel smoothing is best. A natural question, therefore, is under which sample sizes is transformed linear equating preferable to an equipercentile approach.

Figure 29 shows the distribution of differences in error between the two techniques. We ignore scenarios with large differences between groups as, in these cases, we should use alternative techniques anyway. In Figure 29, points below the line indicate that the transformed linear approach is superior on average and points above the line that the smoothed equipercentile approach is preferable. As can be seen, for sample sizes below 500, the transformed linear approach tended to have lower error, whilst for sample sizes greater than 500, the chained equipercentile approach with kernel smoothing had lower error.

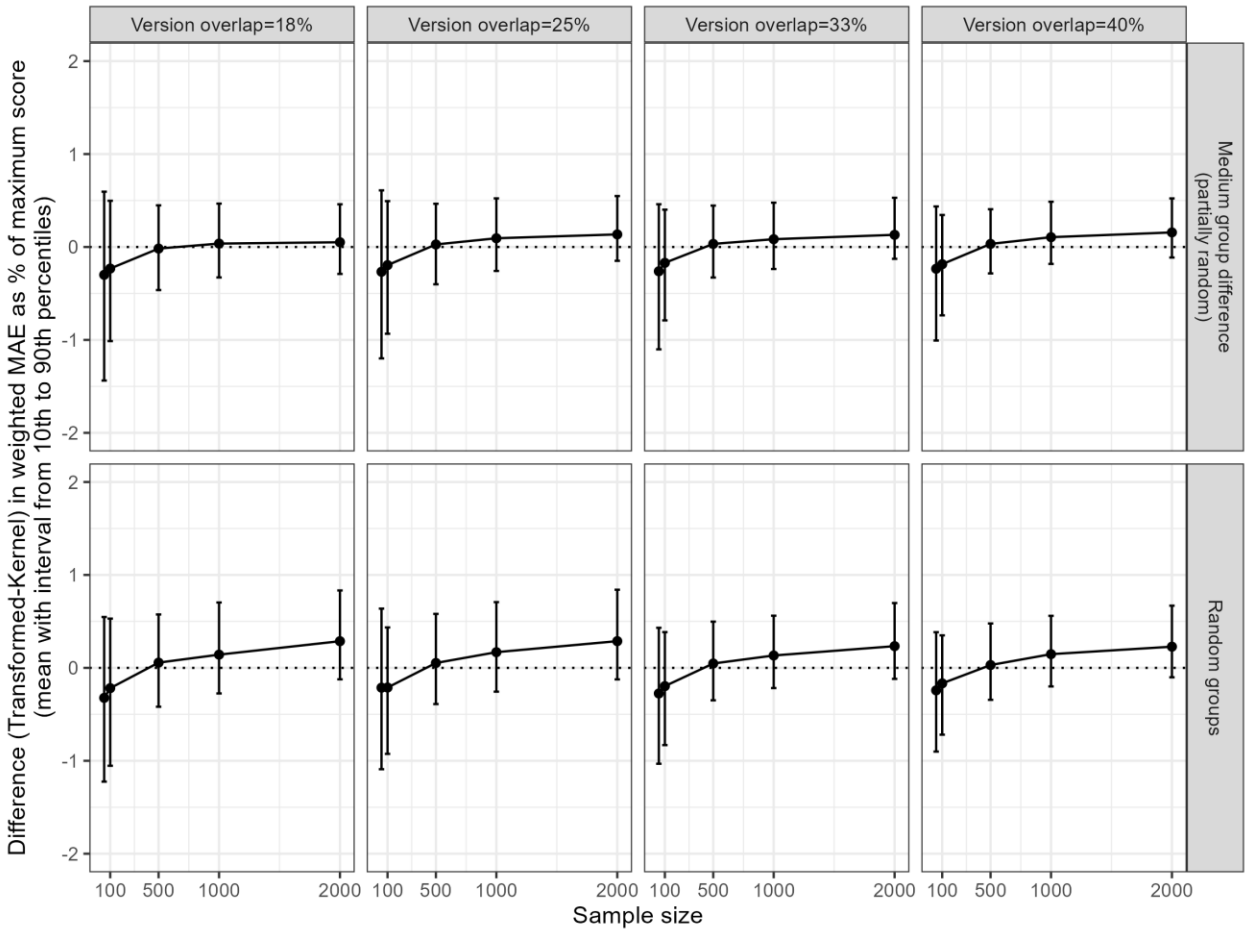

Figure 29. The distribution of differences in weighted MAE between chained linear equating with an initial score transformation and kernel smoothed chained equipercentile equating across a range of scenarios.

Circle-arc vs odds-transform

We now return to the accuracy of circle-arc equating and compare its accuracy to odds-transform equating. We focus upon the Levine method in each case, as it performs relatively well across a range of scenarios, even including those with large differences between groups.

To begin with, Figure 30 displays the distribution of the difference in error between (non-symmetric) circle-arc equating and the odds-transform method. Although the differences in error are some of the smallest anywhere in this study (note that y-axis scale has been shrunk to be between -1 and 1), the odds-transform method is consistently superior (as indicated by the points being above the central horizontal line).

As mentioned earlier, odds-transform equating also has some practical advantages over circle-arc equating (it is guaranteed to be monotonic and is easier to implement). For these reasons we would recommend that the odds-transform approach is always used in favour of non-symmetric circle-arc equating.

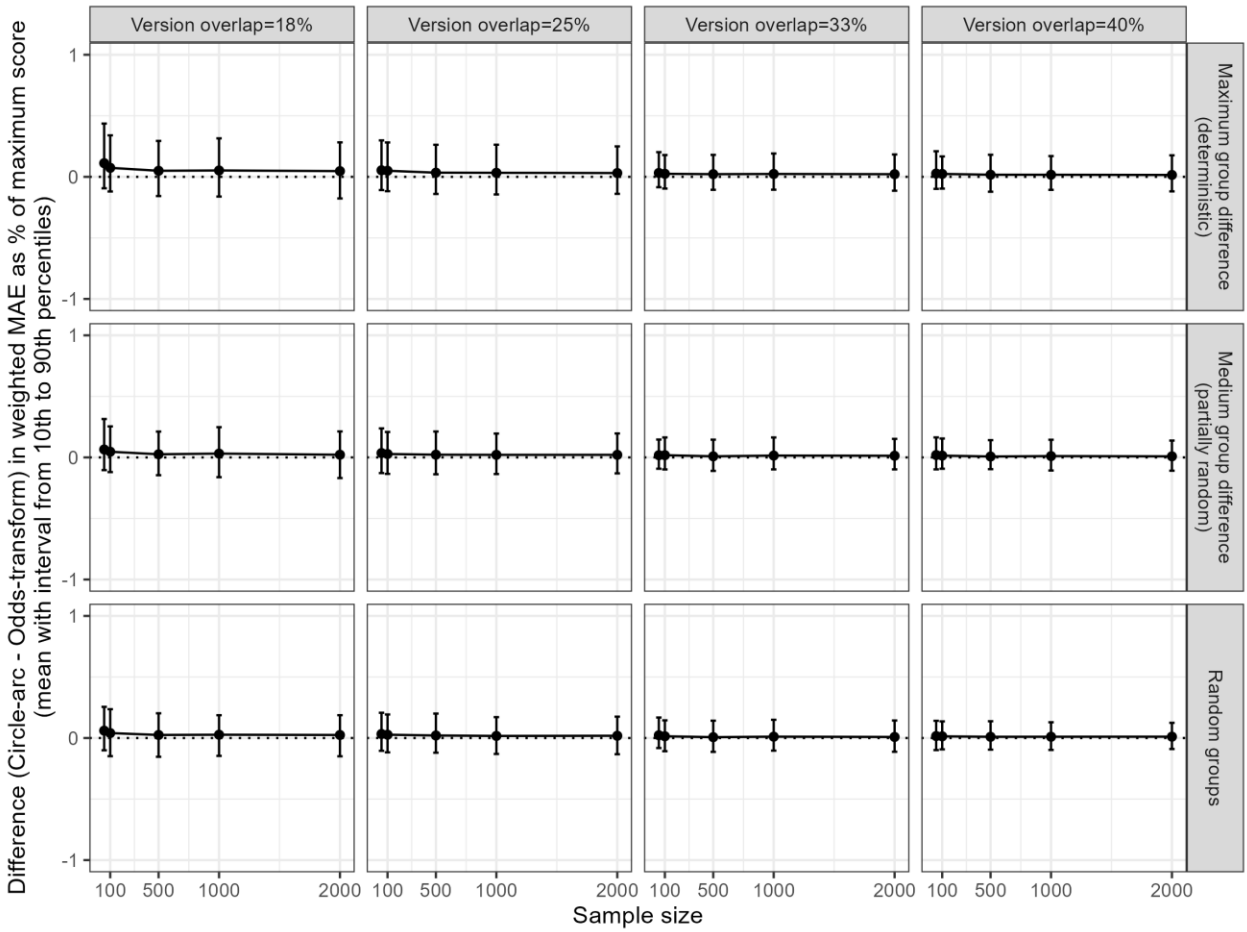

Figure 30. The distribution of differences in weighted MAE between (non-symmetric) circle-arc and odds-transform Levine equating across a range of scenarios.

Figure 31 shows the distribution of differences in weighted MAE between *symmetric* circle-arc and odds-transform methods of Levine equating. The differences here are tiny with even the error bars within the chart now virtually invisible. This indicates that across the scenarios considered in this study, odds-transform and symmetric circle-arc equating are effectively the same thing. However, for practical reasons, such as the guarantee of monotonicity even in extreme circumstances, we would recommend the use of odds-transform equating.

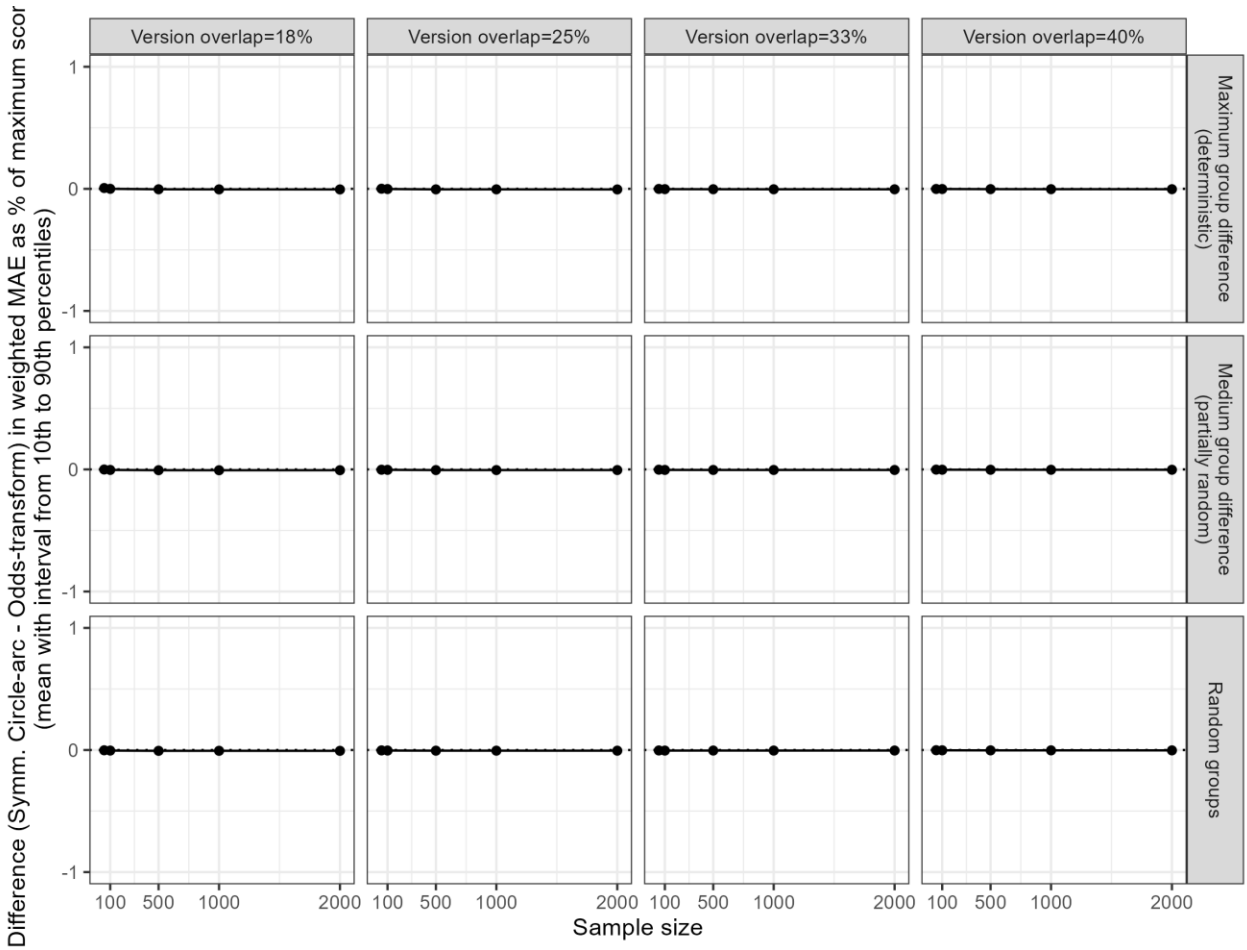

Figure 31. The distribution of differences in weighted MAE between symmetric circle-arc and odds-transform Levine equating across a range of scenarios.

Odds-transform SD vs simple odds-transform

Table 6 shows that, in many scenarios, “SD” variants of odds-transform equating are the most effective. That is, the variant of odds-transform equating that ensures that, at the mean, the gradient of the equating function is equal to the ratio of the standard deviations (within a common synthetic population). This section unpacks this finding a little further.

To being with, Figure 32 shows the distribution of differences in weighted MAE between SD and simple variants of odds-transform Levine equating. Intuitively, we might expect the SD method to require larger sample sizes as it is estimating an extra parameter. This is partially true. However, we can also see that which method is superior is also highly dependent upon the differences between groups. Specifically, where we have large differences between groups, the simple method is superior (all points are above the dotted line).

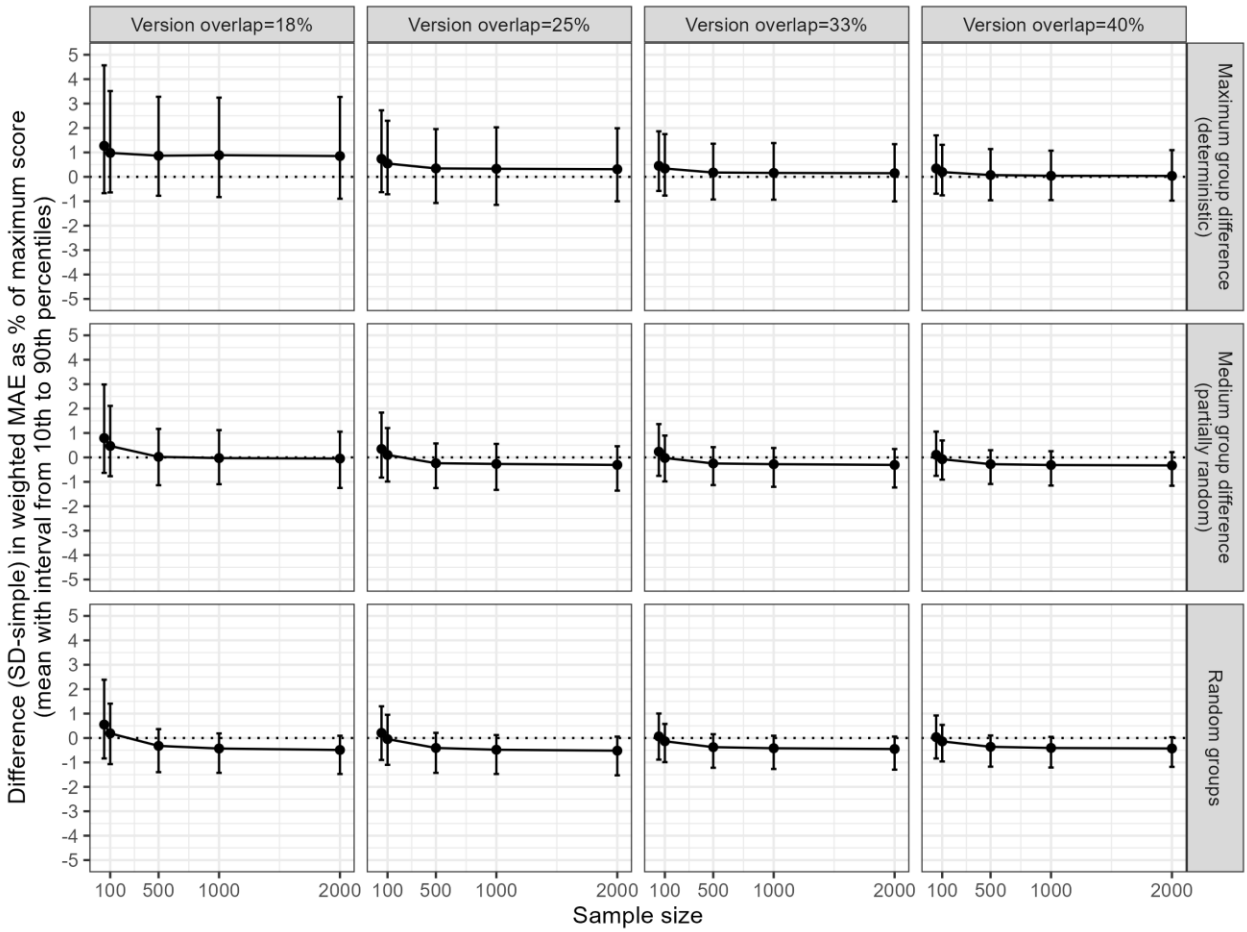

Figure 32. The distribution of differences in weighted MAE between SD and simple (k=1) versions of odds-transform Levine equating across a range of scenarios.

According to Table 6, for all but the largest differences between groups, the chained method is superior to Levine equating. With this in mind, Figure 33 shows the distributions of differences in weighted MAE between SD and simple variants of odds-transform *chained* equating. This figure focusses on scenarios with medium and small differences between groups. As can be seen, and perhaps surprisingly, the SD method appears superior to the simple approach, for all sample sizes. Thus, except for large differences between groups, and perhaps for even smaller sample sizes than those considered in this study, the SD variant of odds-transform chained equating is preferred.

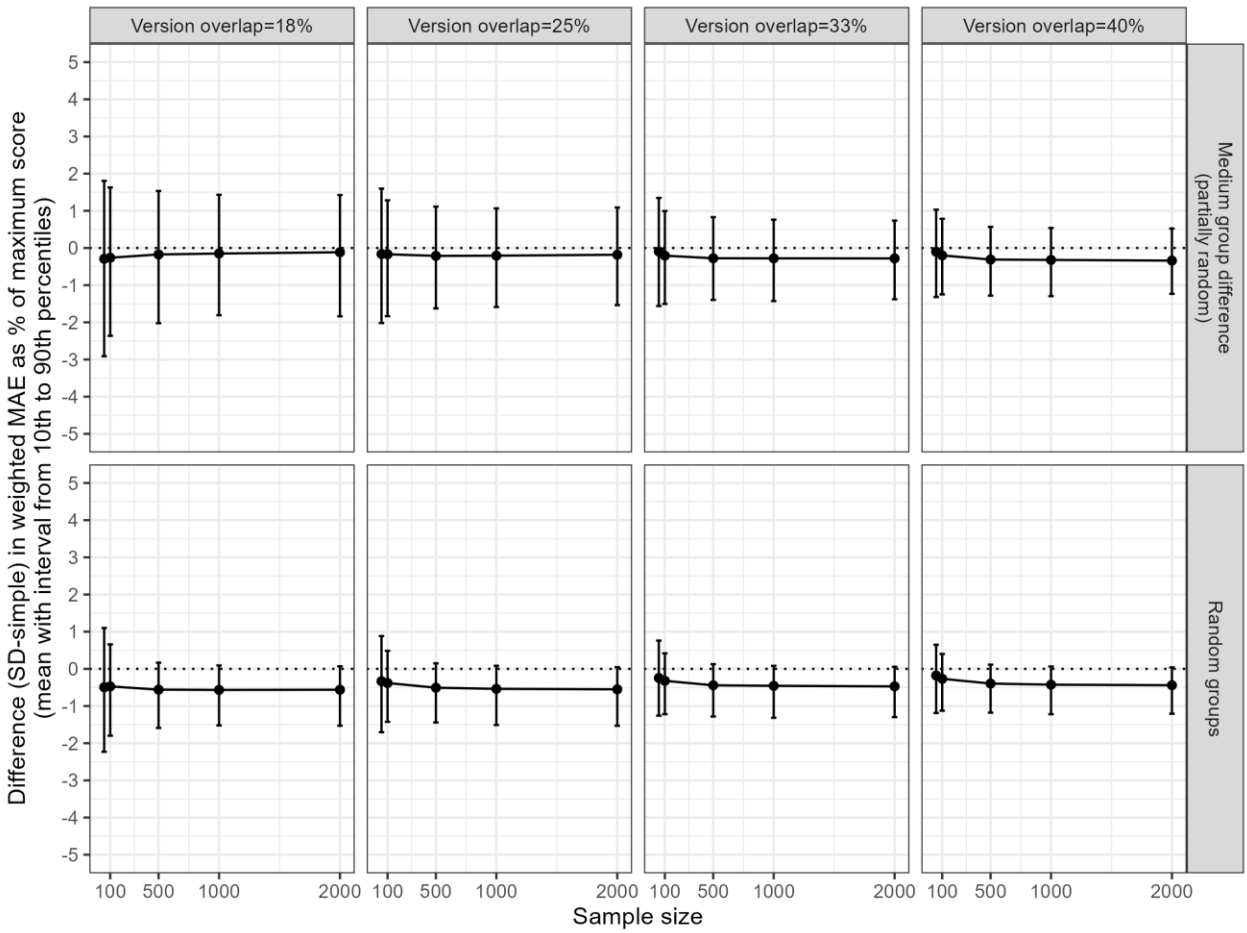

Figure 33. The distribution of differences in weighted MAE between SD and simple (k=1) versions of odds-transform chained equating across a range of scenarios.

Wang-Brennan vs Hybrid Equating

Finally, we compare two modified approaches to frequency estimation equating – the Wang-Brennan (WB) method and the hybrid approach. As described earlier, the hybrid approach builds upon the Levine linear method and can be effective when we have large differences between groups. The Wang-Brennan approach is also specifically designed to be effective in these same scenarios. Figure 34 shows the distribution of differences in error between the two methods across different scenarios. Points below the central horizontal line indicate that the Wang-Brennan approach is superior whereas points above the line indicate that the hybrid approach is superior.

For scenarios with the smallest amount of overlap between test forms, for small and medium group differences, the Wang-Brennan approach is slightly superior in every case. However, for large group differences the hybrid method tends to have lower error for all sample sizes.

In cases with a greater degree of overlap between test forms, the superior method is more variable. However, the hybrid approach tends to be more accurate for small sample sizes whereas the Wang-Brennan method works better for large sample sizes. In short, which of these two modified FE methods is most effective appear somewhat dependent upon the precise scenario.

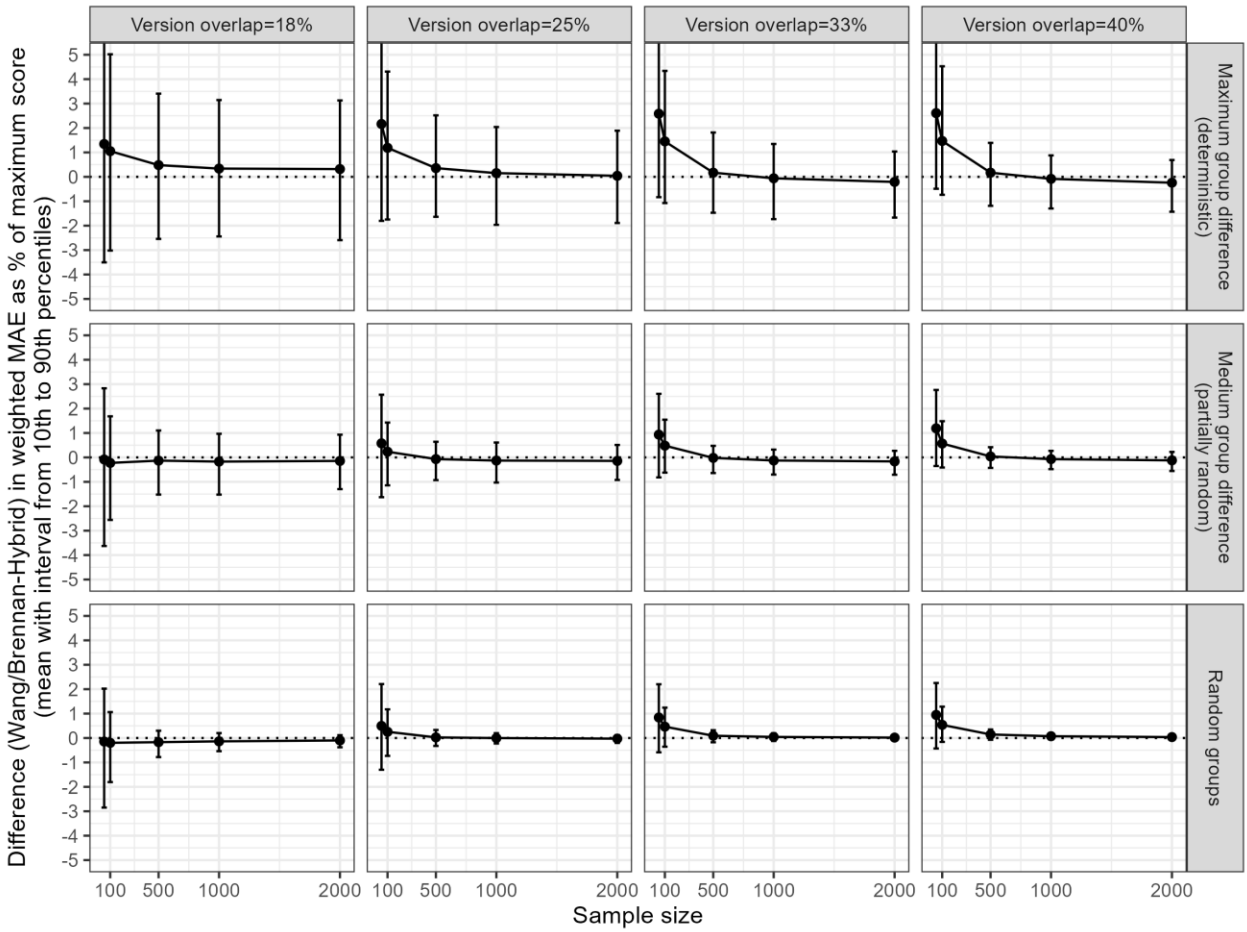

Figure 34. The distribution of differences in weighted MAE between Wang-Brennan and hybrid equating across a range of scenarios.

Summary and discussion

We have undertaken a study that is, to our knowledge, the biggest equating study in the world, ever. The study has tested a wide variety of equating techniques, including essentially all of those suggested in Kolen & Brennan's 2004 book, along with some entirely novel methods, against a very large number of real datasets. In total we have run analyses examining 69 approaches to equating and have completed over 4,000,000 equating analyses[23]. This puts us in a strong position to comment on which techniques are useful and which are not.

Our results have shown us that in benign equating conditions with a reasonably long anchor, fairly large samples, and only minor differences between groups of students taking different forms, there are many equating techniques that will perform adequately. Thus, when the remainder of this discussion labels a method as not useful or says it should be avoided this does not imply the method is bad – only that better alternatives are available.

One practical implication of the analysis presented here is that it allows us to vastly narrow down the number of techniques we need to consider when faced with an equating problem. For many techniques we can identify alternative methods that are either always or nearly always superior regardless of the scenario. So, in the absence of other practical constraints, our results suggest we should:

- Use JML rather than MML or CML estimation of Rasch models.
- Use observed score rather than true score equating for IRT models other than the Rasch model.
- Use concurrent analysis rather than the Stocking-Lord or Haebara methods of metric transformation to align IRT parameters. If circumstances require an approach based upon metric transformation, use the Stocking-Lord rather than the Haebara method.
- Avoid using the unrestricted 3-parameter model for equating unless we have a very strong justification (or, perhaps, if working with purely dichotomous multiple choice test data).
- Avoid using the Tucker method either for linear equating or as a step to calculate the mean test scores in the joint populations to feed into other equating methods. It is only ever marginally superior to a chained approach, and in some scenarios can be substantially worse.
- Use the odds-transform approach to (non-IRT) equating with small samples. It is generally more accurate than the non-symmetric circle-arc method, has fewer quirks and can be extended to work with moderately large sample sizes. Amongst circle-arc methods, the symmetric approach is generally superior to the non-symmetric method. Odds-transform and symmetric circle-arc equating give nearly identical results so the choice between these methods is somewhat a matter of taste.
- If using an equipercentile approach, then smooth the distributions using a kernel approach rather than log-linear pre-smoothing.
- Always apply a simple transformation to help stabilise the conditional standard error of measurement across the score range before applying any form of linear equating.

[23] For each of 60 scenarios and 1,000 datasets in each we have completed 69 equates.

To put this another way, for sample sizes ranging from 50 to 2000, **the following methods of equating should be avoided** in favour of those described above unless there are good reasons for not doing this:

- Identity equating.
- Mean equating.
- Circle-arc equating[24].
- Linear equating (without any initial raw score transformation).
- Observed score equating based upon the Rasch model (true score equating is fine).
- True score equating based upon more complex IRT models (observed score equating is superior).
- Unsmoothed equipercentile equating.
- Equipercentile equating smoothed using either loglinear models or the bump method.
- The Haebara approach to transforming IRT coefficients.

Our results also show us that in situations where we have large differences in ability between groups, then, except for small samples, the use of either IRT models or the Levine method of equating can be quite important and should be preferred if at all possible.

The results also show that whether we use the GPCM or GRM IRT model makes almost no discernible difference to accuracy. Practitioners are free to select either of these perhaps depending upon convenience in terms of software.

One of the most interesting findings was the success of circle-arc and odds-transform equating methods. Such methods were developed for use with small samples, but our analysis has shown that they can be more successful than other classical techniques even when we have a large sample. This is particularly true where we have a large difference between the groups of candidates taking different test forms. The odds-transform method should not, then, only be considered suitable for small samples but also in situations more widely where equating might be considered challenging and there are concerns regarding the use of IRT approaches.

Although it requires some subjective judgement, Table 9 attempts to boil all the analysis in this report down to an essential set of six equating techniques. This summary is based upon earlier analysis in Table 4 and**Error! Reference source not found.**, the broader discussion within the results section, and also by a preference for methods that are easy to implement. Even though under certain circumstances there may be techniques that perform better than those in Table 9, the differences will not be large. In short, Table 9 shows a selection of techniques that are easy to implement and, when used in the suggested circumstances, are unlikely to yield results that are too far from the “truth”. All IRT and Rasch techniques can be implemented using either concurrent estimation or metric transformations depending upon practicalities.

[24] In light of the availability of the odds-transform method.

Table 9: Six of the best equating techniques and when to use them.

Group difference	Sample size		
	50-99	**100-499**	**500 or more**
Less than 0.5 standard deviations	Odds-transform (Chained SD)	Transformed linear (Chained)[25]	Chained equipercentile (Kernel smoothed)
More than 0.5 standard deviations	Odds-transform (Nominal)	Rasch true score equating (JML)	IRT observed score equating (GRM or GPCM)

We would recommend that all practitioners involved in equating analyses should be familiar with the techniques in Table 9 and, ideally, able to implement them.

Note that even those methods are recommended in Table 9 should not be used uncritically. In particular, it is worth remembering that our evaluation is based upon error across the range of scores on form X, but when used in practice, error at *particular scores* may be more important. Hence any practical usage should include some calculation of the standard errors of the technique at the score points at which the method is likely to be used.

Limitations and further work

This research was based upon assessments offered by Cambridge OCR and Cambridge International. Thus, although a diverse range of assessments were included, it clearly does not cover all possible assessment types. In particular, it did not include assessments consisting purely of multiple-choice questions. For the most part, we see no reason why the results should be very different for such tests. However, it is possible that a 3-parameter IRT model may be more effective in such circumstances than for the assessments considered here.

Although this study represents the biggest equating study ever conducted, we have not considered every possible equating scenario. For example, we have not identified the precise transition points in terms of sample size where particular methods become superior to others. Furthermore, we have not explored either *very* small or very large sample sizes, or indeed cases with variable sample sizes. Further work could explore which techniques are most effective in these additional scenarios, and the exact sample sizes at which particular methods become preferable to others.

Similarly, the equating methods included in this study by no means comprise an exhaustive list of all of those available. Further studies in future could evaluate an even wider range of techniques. For example, we have not considered the use of multi-dimensional IRT in equating. Similarly, novel techniques such as the “similar item” approach to equating (Bramley, 2018) could be explored[26]. Nonetheless, although research into the most appropriate equating

[25] The method recommended in this cell of the table was somewhat contentious. The sample size of 100-499 is fairly wide, and the group difference incorporates the random group condition from our analysis, *and* some of the medium difference condition. Many classical techniques work well when group differences are small. Furthermore, if group differences are close to the upper bound in the cell (0.5 standard deviations) a Rasch approach is probably superior. The choice of Transformed linear (chained) equating was made because its average rank in the random or medium group difference conditions when the sample size was either 100 or 500 was 11. No other method had an average rank this high.

[26] Some preliminary attempts to include this technique in our study yielded promising results with the technique potentially being among the best available for medium size differences between groups and large sample sizes. However, further development of the technique is needed to ensure it works efficiently and reliably. This may be a topic for future research.

methods will doubtless continue, hopefully this study has at least reduced the number of techniques that need to be included in future studies of this type.

Finally, it is worth noting that our study has focussed upon situations where we have a genuine anchor test comprising of items in the same subject as the test forms being equated. In many practical scenarios we do not have a formal anchor test of this nature but are instead reliant upon data from other assessments (or other pieces of data) to enable equating (Bramley & Vidal Rodeiro, 2014; Benton & Williamson, 2022; Wallin & Wiberg, 2019). As such, many of the techniques considered in this report are not easily applicable. Having said this, one finding from this research that may be of relevance is the success of odds-transform equating. We have already seen how this technique can be amongst the most effective not only in small samples but more widely in challenging equating circumstances with large differences between groups. Further work could explore whether this technique could also be useful in challenging circumstances where no formal anchor test is available.

Conclusion

There are, perhaps, too many individual findings in this report to draw into a single conclusion. However, it is worth returning to the original purpose of the work: to test a large number of equating methods, under a range of conditions, using a large number of real tests. The aim of this was to provide realistic guidance to those carrying out equating as to which method to use in which situation. We believe that, having carried out “the biggest equating study in the world, ever”, we have achieved this aim, with several methods identified as worthy of consideration in real equating scenarios. This is not to say that methods *not* recommended should *never* be used under any circumstances: our simulated conditions are clearly not exhaustive, so there may well be conditions under which apparently less-favoured methods come into their own. However, we hope that the evidence developed about the methods that *do* work well provides a strong base from which practical decisions can be made when equating has to be carried out for operational or research purposes. Although further work may be required into new methods or performance under particular conditions, we now have a much better understanding of which equating methods to use when which, in turn, should improve the quality of real equating analyses.

References

Albano, A. D. (2015). A general linear method for equating with small samples. *Journal of Educational Measurement*, *52*(1), 55-69.

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

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement*, 72(4), 608-628. https://doi.org/10.1177/0013164411428609.

Benton, T., & Williamson, J. (2022). Which assessment is harder? Some limits of statistical linking. *Research Matters: A Cambridge University Press & Assessment publication*, 34, 26–41. https://www.cambridgeassessment.org.uk/Images/research-matters-34-which-assessment-is-harder.pdf.

Benton, T. (2023). *unimirt: Applications of Unidimensional IRT using the R Package 'mirt'*. R package version 1.0.6. https://github.com/CambridgeAssessmentResearch/unimirt.

Benton, T. (2025). *KernEqWPS: Kernel Equating Without Pre-Smoothing*. R package version 1.0.7. https://github.com/CambridgeAssessmentResearch/KernEqWPS.

Bramley, T. (2018, November). *Evaluating the 'similar items method' for standard maintaining*. Paper presented at the 19th annual conference of the Association for Educational Assessment in Europe, Arnhem-Nijmegen, The Netherlands. https://www.cambridgeassessment.org.uk/Images/515086-evaluating-the-similar-items-method-for-standard-maintaining.pdf.

Bramley, T. & Vidal Rodeiro, C.L. (2014). *Using statistical equating for standard maintaining in GCSEs and A levels*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/182461-using-statistical-equating-for-standard-maintaining-in-gcses-and-a-levels.pdf.

Cai, L., Chung, S. W., & Lee, T. (2023). Incremental model fit assessment in the case of categorical data: Tucker–Lewis index for item response theory modeling. *Prevention Science*, 24(3), 455-466. https://doi.org/10.1007/s11121-021-01253-4.

Chalmers, R. P. (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.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *32*(1), 113-120.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

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

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, 971-988. https://doi.org/10.1214/aos/1176346715.

Hanson, B. A., & Beguin, A. A. (1999). *Separate versus current estimation of IRT item parameters in the common item equating design* (Research Rep. 99–8). Iowa City, IA: ACT, Inc. https://files.eric.ed.gov/fulltext/ED438310.pdf.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. Journal of *Educational Measurement*, 25(2), 97-110. https://doi.org/10.1111/j.1745-3984.1988.tb00295.x.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking (2nd ed.)*. New York: Springer.

Linacre, J.M. (1994). PROX with missing data, or known item or person measures. *Rasch Measurement Transactions*, 8(3), page 378. https://www.rasch.org/rmt/rmt83g.htm.

Livingston, S. A. (2014). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/LIVINGSTON2ed.pdf.

Livingston, S. A., & Kim, S. (2008). *Small-sample equating by the circle-arc method* (Research Report 08-39). Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46(3), 330-343. https://doi.org/10.1111/j.1745-3984.2009.00084.x.

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*, *47*(2), 175-185.

Lord, F. M. (1983). Small N justifies Rasch model. In *New horizons in testing* (pp. 51-61). Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50011-1.

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2025). *dexter: Data Management and Analysis of Tests*. R package version 1.6.0, https://CRAN.R-project.org/package=dexter.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, 11(3), 71-101. https://doi.org/10.1080/15366367.2013.831680.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedogogiske Institut; republished by University of Chicago press, 1980.

Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test Analysis Modules*. R package version 4.2-21. https://CRAN.R-project.org/package=TAM

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, 79(2), 310-334. https://doi.org/10.1177/0013164418783530.

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.

Sun, T., & Kim, S. Y. (2023). Evaluating Equating Methods for Varying Levels of Form Difference. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644231176989.

von Davier, A. A., & Chen, H. (2013). *The Kernel Levine equipercentile observed‐score equating function*. ETS Research Report Series, 2013(2), i-27. https://doi.org/10.1002/j.2333-8504.2013.tb02345.x.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudo-tests constructed from real test data* (Research report 06-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02008.x

Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics*, 44(4), 390-414. https://doi.org/10.3102/1076998619838226.

Wang, T., & Brennan, R. L. (2009). A modified frequency estimation equating method for the common-item nonequivalent groups design. *Applied Psychological Measurement*, 33(2), 118-132. https://doi.org/10.1177/0146621608314607.

Appendix 1: Description of odds-transform equating

Motivation

The odds-transform method of equating is designed along the same lines as circle-arc equating (Livingston and Kim, 2009). The circle-arc method, which is recommended for use in small samples, only requires us to identify three points of equivalence between two tests. Most commonly, these would be the minimum achievable score (i.e., zero), the maximum achievable score on each test, and some estimate of the likely mean score on each test if they were taken by equivalent populations. To apply the method, we calculate the formula for a circle-arc passing through all three points and use this as an estimate of the equating function. According to Livingston and Kim (2010), circle-arc methods are useful because "they require the estimation of only a single point on the equating curve" meaning that they "work well with limited amounts of data" (page 182-183). Furthermore, the approach "results in an estimated equating curve that resembles the curves typically produced by equipercentile equating" (page 183).

In our own simulations, the circle-arc method has been found to work very well. However, the method has certain shortcomings and in rare, but by no means impossible, circumstance can break down entirely. Specifically:

- As noted by Livingston and Kim (2010), one approach to circle-arc equating (sometimes denoted "method 2") is not symmetric and "therefore not truly an equating method" (page 182). A symmetric version is available but requires a more cumbersome set of formulae to implement.
- If the difference in the estimated means of the two forms is large, circle-arc equating can produce equating functions that are not monotonic – that is, as scores on form X increase the equated scores on form Y can sometimes decrease. This issue can occur with both the symmetric and non-symmetric approaches to circle-arc equating.
- If the two test forms have either very different means, or very different maxima, the identified circle-arc may not even pass through the three points identified to begin with. This is because, in these circumstances, the circle associated with the method loops round so that the upper arc passes through some of the three points and the lower arc passes through others. This issue can also occur with both forms of circle-arc equating.

To our knowledge, the last two of the above issues have never been recorded in the equating literature. Odds-transform equating is designed to produce approximately the same equating functions as circle-arc equating, but without the risks identified in the above bullets. Odds-transform equating can also be extended with one extra parameter to allow the method to capture S-shaped equating curves.

Note that, in many typical scenarios where the forms being equated are of similar lengths and have means that are not too different, the results from odds-transform equating will be extremely similar to those from circle-arc equating. As such, odds-transform methods should be seen as minor, but worthwhile, adjustments to an already useful methodology.

Main equating formulae

As with circle-arc equating, to apply odds-transform equating we define three points that we wish the equating relationship to pass through. Typically, these points are labelled (x_1, y_1), (x_2, y_2), and (x_3, y_3). (x_1, y_1) refers to the minimum possible score on each test. Quite often

these are set equal to zero. Alternatively, they are sometimes set to be the scores we might intuitively expect students to achieve through pure guessing. For example, if the tests consist entirely of 4-option multiple-choice tests, then x_1 and y_1 might each be a quarter of the number of items in each test. Similarly, (x_3, y_3) is usually the maximum available score on each test. Finally, (x_2, y_2) is usually the pair of mean scores we'd expect students to achieve on each test in some population.

Odds-transform equating defines raw scores x and y on test forms X and Y as equivalent if the following equation holds.

$$y = y_1 + \frac{(y_3 - y_1)(y_2 - y_1)(x - x_1)^{1/k}(x_3 - x_2)^{1/k}}{(y_3 - y_2)(x_2 - x_1)^{1/k}(x_3 - x)^{1/k} + (y_2 - y_1)(x - x_1)^{1/k}(x_3 - x_2)^{1/k}} \quad (1)$$

In the simplest application of odds-transform equating, the parameter k is simply set to be equal to 1. Alternatively, we can extend the approach to define k so that the gradient of the above equation at the point (x_2, y_2) is equal to the ratio of estimated standard deviations of scores on each form in the define population. To achieve this, we set

$$k = \frac{\sigma_X(x_3 - x_1)(y_2 - y_1)(y_3 - y_2)}{\sigma_Y(y_3 - y_1)(x_2 - x_1)(x_3 - x_2)} \quad (2)$$

Where σ_X and σ_Y are the estimated standard deviations of the scores on form X and form Y in the relevant population.

Simpler expressions can be derived if we assume that x_1 and y_1 are equal to zero. That is, that curve passes through the origin. If we assume this, and replace x_3 and y_3 with m_X and m_Y (denoting the maximum possible scores on forms X and Y), and replace x_2 and y_2 with μ_X and μ_Y (denoting the mean scores on forms X and Y in a common population), then Equation 1 reduces to

$$y = \frac{m_Y \mu_Y x^{1/k}(m_X - \mu_X)^{1/k}}{(m_Y - \mu_Y)\mu_X^{1/k}(m_X - x)^{1/k} + \mu_Y x^{1/k}(m_X - \mu_X)^{1/k}}$$

and Equation 2 reduces to

$$k = \frac{\sigma_X m_X \mu_Y (m_Y - \mu_Y)}{\sigma_Y m_Y \mu_X (m_X - \mu_X)}$$

Derivation

The starting point for this equating method is the PROX approximation for the Rasch ability (θ) of a student based upon the known average difficulty of the test items they have attempted, the variance of item difficulties, test length, and their raw score (Cohen, 1979; Linacre, 1994). Note that, although Equation 1 will be derived starting from the PROX approach, this is purely as a means for identifying a suitable functional form for equating. In particular, it helps us identify a simple functional form that can be used as an alternative to circle-arc equating whilst meeting various important features of equating such as symmetry. It does not imply that that we are

endorsing the use of the PROX method itself for estimating the parameters a Rasch model. Rather, much like circle method itself, the justification for recommending odds-transform equating is "purely practical" (Livingston & Kim, 2009, page 342). However, showing the derivation of Equation 1 helps us understand the links between odds-transform equating and other existing techniques.

For test form X the PROX estimate of θ can be written as:

$$\theta = D_X + \sqrt{(1 + V_X/2.9)} \log\left(\frac{x}{m_X - x}\right) \quad (3)$$

Where D_X is the mean Rasch difficulty of items, V_X is the variance of item difficulties, m_X is the maximum possible score on the test, and x is the student's raw score.

We can create a very similar formula for test form Y.

$$\theta = D_Y + \sqrt{(1 + V_Y/2.9)} \log\left(\frac{y}{m_Y - y}\right) \quad (4)$$

Next, using the principles of true score equating (Kolen & Brennan, 2004, page 176) we say that scores of x and y are equivalent if they lead to the same ability estimate. That is, scores x and y are equivalent if equations 3 and 4 are equal.

$$D_X + \sqrt{(1 + V_X/2.9)} \log\left(\frac{x}{m_X - x}\right) = D_Y + \sqrt{(1 + V_Y/2.9)} \log\left(\frac{y}{m_Y - y}\right) \quad (5)$$

This will be true if and only if

$$\log\left(\frac{x}{m_X - x}\right) = \frac{(D_Y - D_X)}{\sqrt{(1 + V_X/2.9)}} + \frac{\sqrt{(1 + V_Y/2.9)}}{\sqrt{(1 + V_X/2.9)}} \log\left(\frac{y}{m_Y - y}\right) \quad (6)$$

If we define $\log(\beta)$ as $\frac{(D_X - D_Y)}{\sqrt{(1+V_X/2.9)}}$ and k as $\frac{\sqrt{(1+V_Y/2.9)}}{\sqrt{(1+V_X/2.9)}}$ then the above equation can be rewritten as

$$\log\left(\frac{x}{m_X - x}\right) = -\log(\beta) + k \log\left(\frac{y}{m_Y - y}\right) = \log\left(\left(\frac{1}{\beta}\right)\left(\frac{y}{m_Y - y}\right)^k\right) \quad (7)$$

Which will be true if an only if

$$\frac{\beta x}{m_X - x} = \left(\frac{y}{m_Y - y}\right)^k \quad (8)$$

Equation 8 gives a simple equivalence relationship between raw scores on the test forms X and Y with two parameters. All that is required is to estimate sensible values for β and k.
Before exploring how this can be done, note that, like circle-arc equating, and like any form of true score equating, equation 8 already ensures that a score of zero on form X is equivalent to a score of zero on form Y. Furthermore, a score of m_X on form X will be equivalent to a score of m_Y on form Y.

Building on the principles of circle-arc equating we also wish for the mean score on form X (denoted μ_X) to be defined as equivalent to the mean score on form Y (denoted μ_Y). In a random or equivalent groups equating design, μ_X and μ_Y would simply be the sample means of each test form. In the non-equivalent anchor test (NEAT) equating design, μ_X and μ_Y would be the estimated mean score on each test within a common synthetic population. These means might be estimated via a chained procedure (see Livingston & Kim, 2009) or via the nominal weights approach (Babcock, Albano & Raymond, 2012).

From Equation 8 it is easy to see, that μ_X will equaqte to μ_Y if and only if

$$\beta = \left(\frac{m_X - \mu_X}{\mu_X}\right)\left(\frac{\mu_Y}{m_Y - \mu_Y}\right)^k \tag{9}$$

If we substitute this into equation 8 and make y the subject of the equation, this reduces to

$$y = \frac{m_Y \mu_Y x^{1/k}(m_X - \mu_X)^{1/k}}{(m_Y - \mu_Y)\mu_X^{1/k}(m_X - x)^{1/k} + \mu_Y x^{1/k}(m_X - \mu_X)^{1/k}} \tag{10}$$

The above formula will yield a score on form Y equivalent to any given raw score on form X. In usual equating notation, we might write the left-hand side as $e_y(x)$ rather than simply as y.

It is easy to see confirm that Equation 10 is equivalent to Equation 1 if we set $y_1 = x_1 = 0$, $y_2 = \mu_y$, $x_2 = \mu_x$, $y_3 = m_y$, and $x_3 = m_x$. Converting Equation 10 to Equation 1 simply involves imagining that the minimum allowed score on each test is something other than zero. Specifically, this requires us to replace y with $(y - y_1)$, μ_Y with $(y_2 - y_1)$, and m_Y with $(y_3 - y_1)$. Similar changes should be made for x also.

Having derived Equation 10, all that remains now is to determine a sensible value for k. One option is simply to set k equal to 1. This is equivalent to working from Equation 6 and assuming that the item difficulties are equally spread out on each form (i.e., that V_X=V_Y). This approach is the closest in spirit to circle-arc equating in that we have defined a curve (Equation 10) relating x and y that passes through the minimum on each test the maximum available score on each test and the mean score on each test with no further parameters to estimate.

However, rather than simply setting k equal to 1, we may prefer to make an empirical choice for k. We have already ensured that the expression in Equation 10 passes through the minima, maxima and means of each test form. The free choice of k provides the opportunity to ensure that the expression passes through one more point of our choice.

If we define the additional co-ordinates that we wish the equating curve to pass through as (x^*, y^*) then it is easy to derive (starting from equations 8 and 9) that we will need k to be defined as follows.

$$k = \frac{\log\left(\frac{x^*(m_X - \mu_X)}{\mu_X(m_X - x^*)}\right)}{\log\left(\frac{y^*(m_Y - \mu_Y)}{\mu_Y(m_Y - y^*)}\right)} \tag{11}$$

We can use the above equation to enable us to find a value for k such that the gradient of the curve at $x = \mu_x$ will be equal to the ratio of estimated standard deviations of the two test forms. According to Livingston and Kim (2008), a similar idea was suggested by Divgi (1987) in the context of using cubic polynomials to define an equating function.

In our context, to achieve the desired gradient at $x = \mu_x$, we set $x^* = \mu_X + \delta\sigma_X$ and $y^* = \mu_Y + \delta\sigma_Y$ and use Equation 11. This approach means that, each score on form X will equate to a score on form X which is the same number of standard deviations (δ) from the mean. This implies that if we solve Equation 11 as $\delta \to 0$ then the gradient at the mean will be equal to $\frac{\sigma_Y}{\sigma_X}$. As such we must solve the following equation,

$$k = \lim_{\delta \to 0} \frac{\log\left(\frac{(\mu_X + \delta\sigma_X)(m_X - \mu_X)}{\mu_X(m_X - \mu_X + \delta\sigma_X)}\right)}{\log\left(\frac{(\mu_Y + \delta\sigma_Y)(m_Y - \mu_Y)}{\mu_Y(m_Y - \mu_Y + \delta\sigma_Y)}\right)} \tag{12}$$

To solve this, first, we rearrange the numerator as follows

$$\log\left(\frac{(\mu_X + \delta\sigma_X)(m_X - \mu_X)}{\mu_X(m_X - \mu_X + \delta\sigma_X)}\right) = \log(m_X - \mu_X) + \log(\mu_X + \delta\sigma_X) - \log(\mu_X) - \log\left(m_X - \mu_X + \delta\sigma_X\right)$$

By using a first order Taylor expansion, we know that, if δ is small, then we can approximate $\log(\mu_X + \delta\sigma_X)$ by $\log(\mu_X) + \frac{\delta\sigma_X}{\mu_X}$. Similarly, we can approximate $\log\left(m_X - \mu_X + \delta\sigma_X\right)$ by $\log(m_X - \mu_X) + \frac{\delta\sigma_X}{m_X - \mu_X}$.

When we substitute these expressions into the numerator of Equation 12 it reduces to.

$$\frac{\delta\sigma_X}{\mu_X} + \frac{\delta\sigma_X}{m_X - \mu_X} = \frac{\delta\sigma_X m_X}{\mu_X(m_X - \mu_X)}$$

An essentially equivalent approximation can be made for the denominator in Equation 12. Thus, as $\delta \to 0$ we find that k can be calculated as

$$k = \lim_{\delta \to 0} \frac{\log\left(\frac{(\mu_X + \delta\sigma_X)(m_X - \mu_X)}{\mu_X(m_X - \mu_X + \delta\sigma_X)}\right)}{\log\left(\frac{(\mu_Y + \delta\sigma_Y)(m_Y - \mu_Y)}{\mu_Y(m_Y - \mu_Y + \delta\sigma_Y)}\right)} = \lim_{\delta \to 0} \frac{\left(\frac{\delta\sigma_X m_X}{\mu_X(m_X - \mu_X)}\right)}{\left(\frac{\delta\sigma_Y m_Y}{\mu_Y(m_Y - \mu_Y)}\right)} = \frac{\left(\frac{\sigma_X m_X}{\mu_X(m_X - \mu_X)}\right)}{\left(\frac{\sigma_Y m_Y}{\mu_Y(m_Y - \mu_Y)}\right)}$$

This provides an alternative approach to estimating k. Specifically, simplifying the equation above, we compute k via the following expression.

$$k = \frac{\sigma_X m_X \mu_Y (m_Y - \mu_Y)}{\sigma_Y m_Y \mu_X (m_X - \mu_X)} \tag{13}$$

Equation 13 is equivalent to Equation 2 if we make the substitutions allowing for the minimum available score on each test being something other than zero as before.

Comparing odds-transform and circle-arc equating functions for test forms with equal maxima

To give some insight into the differences between odds-transform and circle-arc equating, Figure 35 compares these equating functions for two test forms with maximum scores of 10. For the purposes of this figure, we set k equal to 1 for odds-transform equating. Both forms of circle-arc equating are shown. The different panels show all possible equating functions where the mean on form X (that is, x_2) is any of 4, 5 or 6, and the mean on form Y (that is, y_2) is any of 5, 6, or 7. In every case $x_1 = y_1 = 0$. As can be seen, where the means on the two forms differ by no more than 1 score point out of 10 (e.g., the diagonal panels, from top left to bottom right), the three equating methods are almost indistinguishable. At differences of 2 score points out of 10 (the two panels adjacent to the bottom left-hand corner one), the non-symmetric version of circle-arc equating starts to provide noticeably different curves to the other two methods (which remain indistinguishable). At differences of 3 score points between the means (the bottom left-hand panel), serious issues begin to emerge with the non-symmetric version of circle-arc equating in that the equating function is not monotonic and equates scores of 8 or 9 on form X to scores above 10 (i.e., outside the allowable range) on form Y. However, the odds-transform and symmetric circle-arc methods continue to produce very similar results.

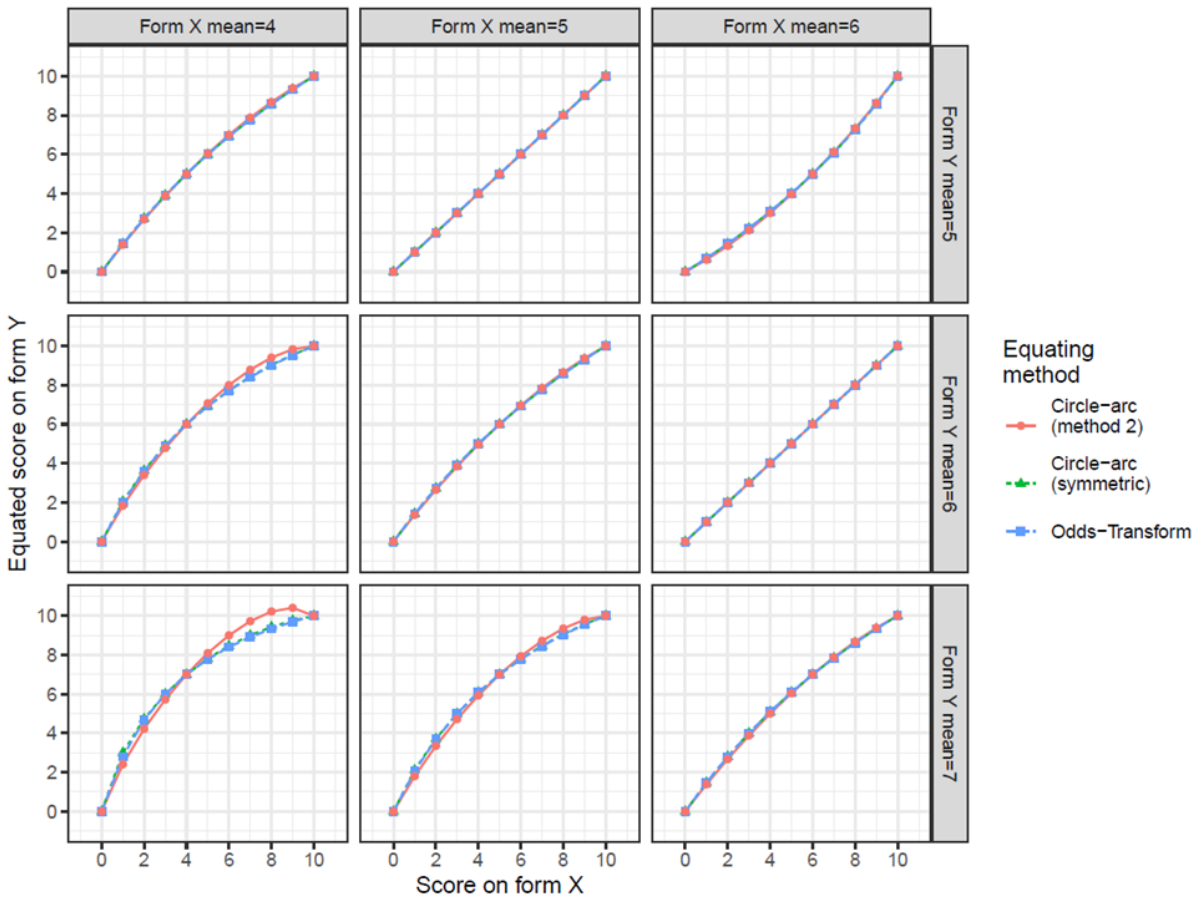

Figure 35: Odds-transform and circle-arc equating functions for test forms with a maximum of 10 and reasonably similar means.

Figure 36 illustrates what happens if the differences in means become even more extreme. In this chart, the panels show all possible equating functions where the mean on form X (that is, x_2) is any of 1, 2 or 3, and the mean on form Y (that is, y_2) is any of 7, 8, or 9. In every one of these cases, non-symmetric circle-arc equating yields a non-monotonic equating function that

goes well beyond the allowable score range for form Y. There is also some evidence of problems with the symmetric method. A number of panels show the method yielding a non-monotonic equating function. Furthermore, in every panel except the one in the top right-hand corner, a score of 0 on form X is not mapped to a score of 0 on form Y. This because the identified circle formula would only pass through (0,0) as part of its lower arc whereas the upper arc passes through the remaining two points. In contrast, for all values, the odds-transform method continues to yield monotonic equating functions and passes through each of the three co-ordinates targeted by the method.

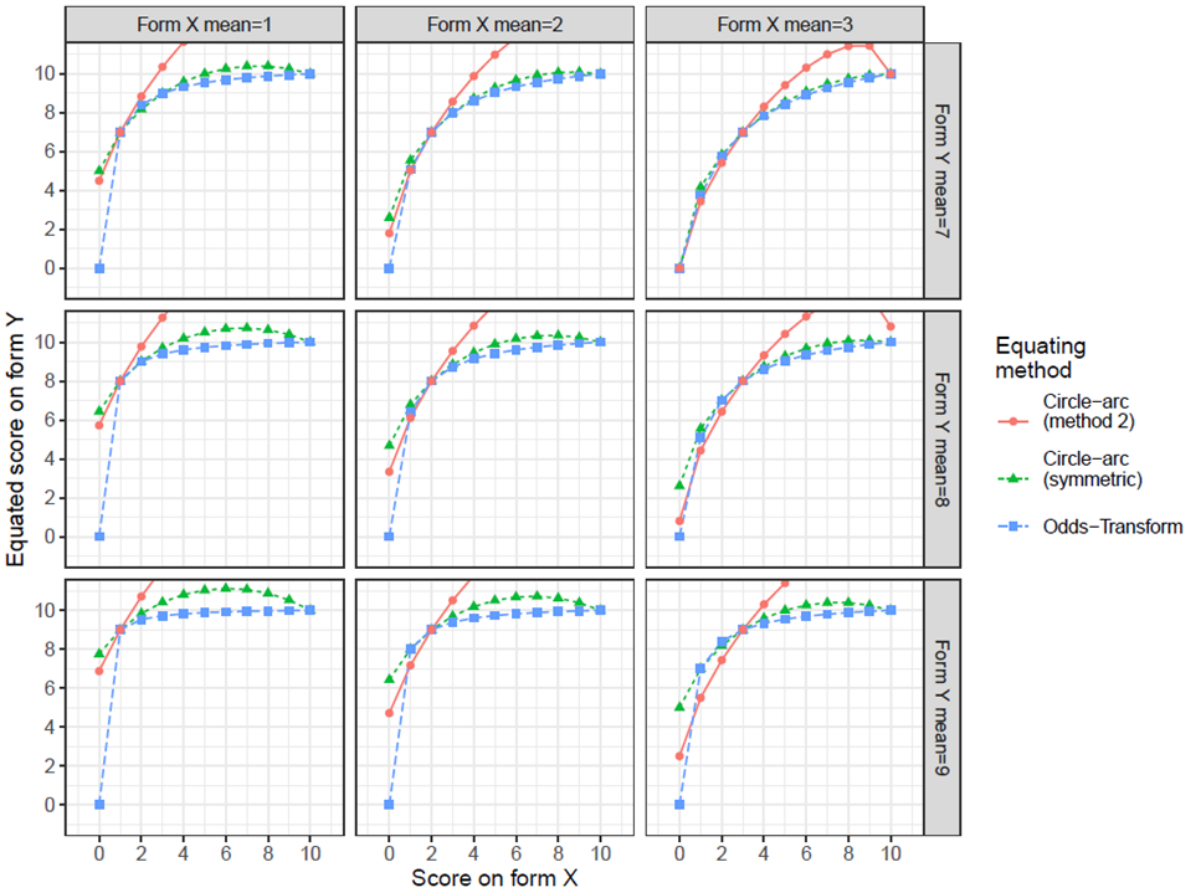

Figure 36: Odds-transform and circle-arc equating functions for test forms with a maximum of 10 and very different means.

Undoubtably, the problems with circle-arc equating identified in this section are rare in practice. If we are equating two test forms, we would generally hope they are of reasonably similar difficulty, and so, in a common population, their means would be similar. In these scenarios, circle-arc equating, and odds-transform equating will give very similar results.

However, whilst rare, the problems with circle-arc equating identified in this section can and do occur in practice. The scenarios where they occur may, perhaps, be described as linking rather than equating as they involve identifying equivalent scores on test forms created to very different specifications (perhaps by different assessment organisations). As such, equated scores on the different test versions are unlikely to indicate interchangeable information about students. Nonetheless, when faced with such scenarios, the educational measurement professional may be required to give their best estimates of score equivalencies and therefore needs access to tools that will not break down.

Comparing odds-transform and circle-arc equating functions for test forms with unequal maxima

Similar issues to those above, may occur if the lengths of the tests being equated (or more likely linked) are very different, even if the means as a proportion of the available maximum are fairly similar. This issue might be particularly problematic where we wish to implement chained versions of circle-arc equating in the non-equivalent anchor test (NEAT) design where the anchor test is much shorter than either of the test forms being equated (see, Albano, 2015, for an example where chained circle-arc equating was trialled).

To illustrate the problem, Figure 37 compares the various equating functions for two test forms with maximum scores of 10 and 50 respectively. Again, we set k equal to 1 for odds-transform equating for this figure. The different panels show all possible equating functions where the mean on form X (that is, x_2) is any of 4, 5 or 6 out of 10, and the mean on form Y (that is, y_2) is any of 25, 30, or 35 out of 50. In every case $x_1 = y_1 = 0$. Where the means on the two forms are the same as a proportion of the available maximum (e.g., 5 out of 10 and 25 out of 50) the different methods yield the same answers. However, where there are large differences in the proportional means (the three panels closest to the bottom left-hand corner) problems are seen with both forms of circle-arc equating. For example, if the form X mean is 4 out 10 and the form Y mean is 35 out of 50, neither circle-arc method passes through (0,0) and only the symmetric method passes through (10,50). This issue is again caused by upper arc of the identified circles passing through some of the desired points whilst the lower arc passes through others.

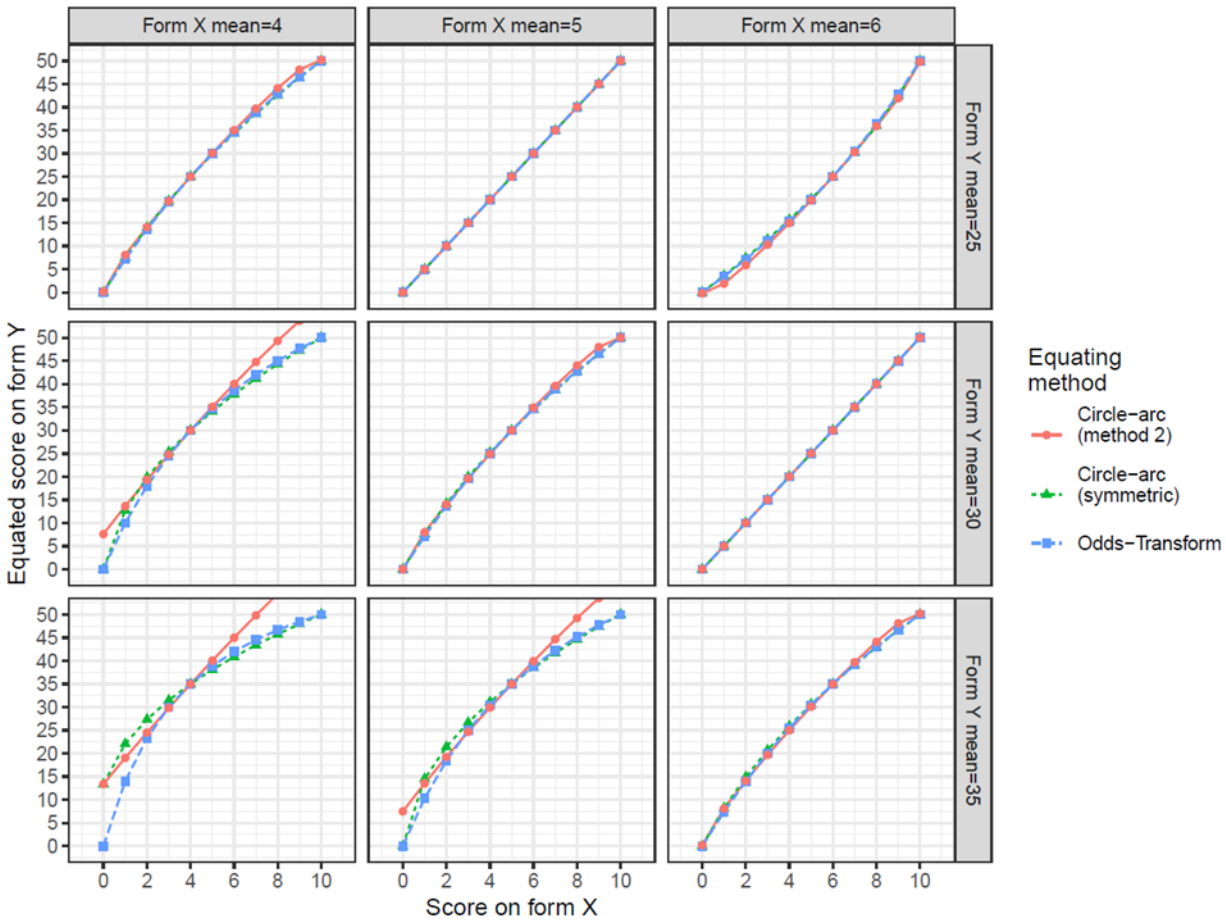

Figure 37: Odds-transform and circle-arc equating functions for test forms with maxima of 10 and 50 respectively and various means.

The particular issues in Figure 37 can, of course, be resolved by rescaling both test forms to have a common maximum. However, this indicates another difference between circle-arc and odds-transform equating – that of scale invariance. Specifically, if $x_1 = y_1 = 0$, odds-transform equating is invariant to multiplicative transformations. For example, imagine the method maps a score of 10 on form X to a score of 15 on form Y. Next imagine that scores are transformed so that all form X scores are multiplied by 2 and all form Y scores are multiplied by 3. We can be sure that re-applying odds-transform equating to the transformed scores will map a (transformed) score of 20 on form X to a (transformed) score of 45 on form Y. This type of invariance is not a property of either approach to circle-arc equating. In either case, a change of scale changes the shape of the equating function.

Appendix 2: Brief descriptions of all equating methods used in analysis

This appendix provides a few sentences describing the basic idea of each of the equating methods used in analysis. More detailed descriptions can usually be found in Kolen & Brennan (2004). Where the method is not described in Kolen & Brennan's book, an explicit reference where further details can be found has been provided. Methods are listed in alphabetical order.

For the purposes of our research the synthetic population always refers to joint population of candidates across both test forms. That is, we are interested in the anchor score distribution across all sets of candidates in the analysis combined (i.e., those who did form X and those who did form Y).

- **Chained Equipercentile (Bump smoothed)**. Chained equipercentile equating but with a single candidate with a weight of one millionth added to each available integer score to help address any possible gaps.
- **Chained Equipercentile (Kernel smoothed)**. Chained equipercentile equating but with kernel pre-smoothing applied to all distributions (Benton, 2024).
- **Chained Equipercentile (Loglinear smoothed)**. Chained equipercentile equating but with a log linear model used to pre-smooth all score distributions.
- **Chained Equipercentile (Unsmoothed)**. Chained equipercentile equating without any pre-smoothing applied.
- **Circle-arc (Braun-Holland)**. The Braun-Holland method weights the data on each form so that the full distribution of anchor test scores matches that in the synthetic population (the same process as for Frequency Estimation or "FE" equating). The mean score on each test form within the weighted data is used as an input to non-symmetric circle-arc equating.
- **Circle-arc (Chained)**. Chained (i.e., successive) circle-arc equating of Form X to the anchor and the anchor to Form Y.
- **Circle-arc (Levine)**. The Levine Observed score method is an attempt to use regression to estimate the mean score (and standard deviation) on each form in the synthetic population whilst accounting for the impact of measurement error in the anchor. To do this, the reliability of the anchor is estimated purely based on test length (i.e. maximum available score) relative to the form X and Y tests as well as the correlation with each test form. Having used the Levine method to estimate means, these are used within the non-symmetric circle-arc approach to equating.
- **Circle-arc (Nominal)**. The nominal method uses simple assumption that the proportion of the maximum score achieved on each test should change between each group and the synthetic population by the same extent as the anchor test (see Babcock et al, 2012). Once this approach is used to estimate means these are used within the non-symmetric circle-arc approach to equating.
- **Circle-arc (Tucker)**. The Tucker method builds upon the usual assumptions of linear regression to estimate the mean score (and standard deviation) on each form in the synthetic population. Having used the Tucker method to estimate means, these are used within the non-symmetric circle-arc approach to equating.
- **FE Equipercentile (Bump smoothed)**. Frequency estimation (FE) equipercentile equating weights the data on each form so that the full distribution of anchor test scores matches that in the synthetic population. Then, assuming that this has made the two

groups equivalent, equipercentile equating of the Form X and Form Y score distributions is completed. A single candidate with a weight of one millionth is added to each available integer score to help address any possible gaps.

- **FE Equipercentile (Kernel smoothed)**. This method weights the data on each form so that first four moments of the anchor test score distribution matches that in the synthetic population. Then, assuming that this has made the two groups equivalent, equipercentile equating of the Form X and Form Y score distributions is completed. Kernel pre-smoothing applied to the score distributions in the (weighted) equivalent groups (Benton, 2024).
- **FE Equipercentile (Loglin smoothed)**. FE equipercentile equating but with log linear pre-smoothing applied to the score distributions in the equivalent groups (Benton, 2024).
- **FE Equipercentile (Modified kernel hybrid)**. A hybrid of the Levine observed score linear method (to account for measurement error in the anchor) and smoothed FE equipercentile equating (to account for non-linear equating relationships). See von Davier & Chen (2013) for further details.
- **FE Equipercentile (Modified WB unsmoothed)**. A modified frequency estimation approach that attempts to account for measurement error in the anchor test. This is done in two stages. Firstly, Cronbach's alpha is used to estimate reliability in each test. Then, using this, anchor scores in one test form are adjusted so that group differences are expanded to the extent that would be seen if the anchor test had no measurement error. After this, unsmoothed FE equipercentile equating is applied in the usual way. See Wang & Brennan (2009) for further details.
- **FE Equipercentile (Unsmoothed)**. FE equipercentile equating with no pre-smoothing applied.
- **GPCM (MML Concurrent Observed)**. Using all data from both test forms combined, fit a generalised partial credit model (GPCM). This done using a multiple-group IRT estimation based upon marginal maximum likelihood (MML). Item parameters are defined to be equal across groups. In one group the mean ability is defined to be zero and the standard deviation to be equal to 1. In the other, both the mean and the standard deviation of ability is estimated. Normal ability distributions are assumed in each group. After estimation, the Lord-Wingersky method is used to estimate the score distribution of each form in the synthetic population and equipercentile equating is applied to these.
- **GPCM (MML Concurrent True)**. Model fitting as above. True score equating implies that raw test scores are deemed as equivalent if they are associated with the same level of ability in their test characteristic curves.
- **GPCM (MML Haebara Observed)**. GPCM model fitted to data from each form separately. In each case a normal ability distribution with a mean of zero and a standard deviation of one is assumed. After fitting separately models, item parameters from form Y are transformed to be on the same scale as those from form X using the Haebara method (essentially this ensures that the item characteristic curves of anchor test items are as similar as possible). Having done this, the Lord-Wingersky method is used to estimate the score distribution of each form in the synthetic population and equipercentile equating is applied to these.
- **GPCM (MML Haebara True)**. Model fitting and metric transformation as above. True score equating implies that raw test scores are deemed as equivalent if they are associated with the same level of ability in their test characteristic curves.

- **GPCM (MML SL Observed)**. Same method as GPCM (MML Haebara Observed) but the Stocking-Lord method is used to identify the metric transformation rather than the Haebara method. Essentially this ensures that the test characteristic curves for the subsections of each form that comprise the anchor test are as similar as possible.
- **GPCM (MML SL True)**. Same method as GPCM (MML Haebara True) but the Stocking-Lord method is used to identify the metric transformation rather than the Haebara method.
- **GRM (MML Concurrent Observed)**. Same method as GPCM (MML Concurrent Observed) but using the graded response model (GRM) rather than the GPCM.
- **GRM (MML Concurrent True)**. Same method as GPCM (MML Concurrent True) but using the graded response model (GRM) rather than the GPCM.
- **GRM (MML Haebara Observed)**. Same method as GPCM (MML Haebara Observed) but using the graded response model (GRM) rather than the GPCM.
- **GRM (MML Haebara True)**. Same method as GPCM (MML Haebara True) but using the graded response model (GRM) rather than the GPCM.
- **GRM (MML SL Observed)**. Same method as GPCM (MML SL Observed) but using the graded response model (GRM) rather than the GPCM.
- **GRM (MML SL True)**. Same method as GPCM (MML SL True) but using the graded response model (GRM) rather than the GPCM.
- **GRM/3PL (MML Haebara Observed)**. Same method as GPCM (MML Concurrent Observed) but using the graded response model (GRM) for polytomous items and the 3-parameter model for dichotomous items rather than the GPCM.
- **GRM (MML Haebara True)**. Same method as GPCM (MML Haebara True) but using the graded response model (GRM) for polytomous items and the 3-parameter model for dichotomous items rather than the GPCM.
- **GRM (MML SL Observed)**. Same method as GPCM (MML SL Observed) but using the graded response model (GRM) for polytomous items and the 3-parameter model for dichotomous items rather than the GPCM.
- **GRM (MML SL True)**. Same method as GPCM (MML SL True) but using the graded response model (GRM) for polytomous items and the 3-parameter model for dichotomous items rather than the GPCM.
- **Identity**. Simply assume that scores that represent the same proportion of the maximum available score on each form are equivalent.
- **Linear (Braun-Holland)**. The Braun-Holland method weights the data on each form so that the full distribution of anchor test scores matches that in the synthetic population (the same process as for Frequency Estimation of “FE” equating). The means and standard deviations of scores on each test form within the weighted data are used as an input to linear equating.
- **Linear (Chained)**. Chained (i.e., successive) linear equating of Form X to the anchor and the anchor to Form Y.
- **Linear (Levine Observed)**. The Levine Observed score method is an attempt to use regression to estimate the means and standard deviations of each test form in the synthetic population whilst accounting for the impact of measurement error in the anchor. To do this, the reliability of the anchor is estimated purely based on test length (i.e. maximum available score) relative to the form X and Y tests as well as the correlation with each test form. The estimated means and standard deviations of scores on each test form within the weighted data are used as an input to linear equating.

- **Linear (Levine True)**. As above but estimates the means and standard deviations of true scores on each form in the synthetic population. These estimated means and standard deviations are used in the usual linear equating.
- **Linear (PSE)**. This method weights the data on each form so that first four moments of the anchor test score distribution matches that in the synthetic population. The means and standard deviations of scores on each test form within the weighted data are used as an input to linear equating.
- **Linear (Tucker)**. The Tucker method builds upon the usual assumptions of linear regression to estimate the means and standard deviations of scores on each test form in the synthetic population. These estimated means and standard deviations are used in the usual linear equating.
- **Mean (Braun-Holland)**. Same method as Circle-arc (Braun-Holland) but estimated means feed into mean equating rather than a circle-arc method.
- **Mean (Chained)**. First mean equate form X to the anchor. This means first calculating the different in the mean proportion of the maximum mark achieved on each form. Call this difference D. If a proportional score on form X is equal to a proportional score on the anchor plus *D*, the two scores are deemed equivalent. Use the same approach to equate the anchor to form Y and form a chained equate.
- **Mean (Levine)**. Same method as Circle-arc (Levine) but estimated means feed into mean equating rather than a circle-arc method.
- **Mean (Nominal)**. Same method as Circle-arc (Nominal) but estimated means feed into mean equating rather than a circle-arc method.
- **Mean (Tucker)**. Same method as Circle-arc (Tucker) but estimated means feed into mean equating rather than a circle-arc method.
- **Odds-transform (Chained SD)**. Chained (i.e., successive) odds-transform equating of Form X to the anchor and the anchor to Form Y. The odds-transform approach that tries to ensure that, not only are mean scores on each form deemed equivalent, but that scores one standard deviation from the mean are also deemed equivalent (see Appendix 1).
- **Odds-transform (Chained)**. As above with the k parameter set equal to 1 (see Appendix 1).
- **Odds-transform (Levine SD)**. The Levine approach to estimating the means and standard deviation of test scores (see Linear (Levine Observed)). Estimated means and standard deviations are used in the odds-transform approach to equating (see Appendix 1).
- **Odds-transform (Levine)**. As above with the k parameter set equal to 1 (see Appendix 1).
- **Odds-transform (Nominal)**. Same as circle-arc (nominal) but the estimated means feed into odds-transform approach to equating rather than a circle-arc approach.
- **Odds-transform (Tucker SD)**. The Tucker approach to estimating the means and standard deviation of test scores (see Linear (Tucker)). Estimated means and standard deviations are used in the odds-transform approach to equating (see Appendix 1).
- **Odds-transform (Tucker)**. As above with the k parameter set equal to 1 (see Appendix 1).
- **PCM (CML Haebara True)**. Same method as GPCM (MML Haebara True) but using the Rasch partial credit model (PCM) and with item parameter (difficulty) estimation completed using conditional maximum likelihood (CML).

- **PCM (CML SL True)**. Same method as GPCM (MML SL True) but using the Rasch partial credit model (PCM) and with item parameter (difficulty) estimation completed using conditional maximum likelihood (CML).
- **PCM (CML Concurrent True)**. Same method as GPCM (MML Concurrent True) but using the Rasch partial credit model (PCM) and with item parameter (difficulty) estimation completed using conditional maximum likelihood (CML).
- **PCM (JML Haebara True)**. Same method as GPCM (MML Haebara True) but using the Rasch partial credit model (PCM) and with item parameter (difficulty) estimation completed using joint maximum likelihood (JML).
- **PCM (JML SL True)**. Same method as GPCM (MML SL True) but using the Rasch partial credit model (PCM) and with item parameter (difficulty) estimation completed using joint maximum likelihood (JML).
- **PCM (JML Concurrent True)**. Same method as GPCM (MML Concurrent True) but using the Rasch partial credit model (PCM) and with item parameter (difficulty) estimation completed using conditional maximum likelihood (JML).
- **PCM (MML Concurrent Observed)**. Same method as GPCM (MML Concurrent Observed) but using the Rasch partial credit model (PCM).
- **PCM (MML Concurrent True)**. Same method as GPCM (MML Concurrent True) but using the Rasch partial credit model (PCM).
- **PCM (MML Haebara Observed)**. Same method as GPCM (MML Haebara Observed) but using the Rasch partial credit model (PCM).
- **PCM (MML Haebara True)**. Same method as GPCM (MML Haebara True) but using the Rasch partial credit model (PCM).
- **PCM (MML SL Observed)**. Same method as GPCM (MML SL Observed) but using the Rasch partial credit model (PCM).
- **PCM (MML SL True)**. Same method as GPCM (MML SL True) but using the Rasch partial credit model (PCM).
- **Symmetric circle-arc (Braun-Holland)**. Same method as circle-arc (Braun-Holland) but relying on the symmetric circle-arc method of equating.
- **Symmetric circle-arc (Chained)**. Same method as circle-arc (chained) but relying on the symmetric circle-arc method of equating.
- **Symmetric circle-arc (Levine)**. Same method as circle-arc (Levine) but relying on the symmetric circle-arc method of equating.
- **Symmetric circle-arc (Nominal)**. Same method as circle-arc (nominal) but relying on the symmetric circle-arc method of equating.
- **Symmetric circle-arc (Tucker)**. Same method as circle-arc (Tucker) but relying on the symmetric circle-arc method of equating.
- **Transformed Linear (Chained)**. Same method as Linear (Chained) but with scores on all test forms initially transformed to stabilise the conditional standard error of measurement. After equating this transformation is reversed to put all scores back on the raw score scale (see report section entitled “Results part 2: alternative methods for small samples and small differences between groups”).
- **Transformed Linear (Tucker)**. Same method as Linear (Tucker) but with scores on all test forms initially transformed to stabilise the conditional standard error of measurement. After equating this transformation is reversed to put all scores back on the raw score scale (see report section entitled “Results part 2: alternative methods for small samples and small differences between groups”).

Appendix 3: Supplementary results

In this appendix, we present extra results from the main simulation.

First, we present the "best methods" results according to RMSE, mirroring the results presented in Table 4 (considering all methods) and Table 6**Error! Reference source not found.** (considering only classical equating methods). After this, we present comparisons of all methods under different sample sizes for the maximum group difference and random groups conditions, mirroring results in Figure 3 to Figure 7; note that we only present these results for the 25% overlap condition, given the modest impact of anchor length on outcomes.

Table 10. The best-performing equating methods (i.e., those with the smallest mean weighted RMSE) under each simulation scenario.

Group difference	Form overlap	Sample size				
		50	100	500	1000	2000
Maximum	18%	Symmetric circle-arc (Nominal)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	GPCM (MML Conc Obs)	GRM (MML Conc Obs)
	25%	PCM (JML Conc True)	PCM (JML Conc True)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
	33%	PCM (JML Conc True)	PCM (JML Conc True)	GRM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
	40%	Symmetric circle-arc (Levine)	PCM (JML Conc True)	GRM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
Medium	18%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
	25%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)	GPCM (MML Conc Obs)
	33%	Odds-transform (Chained SD)	PCM (JML Conc True)	GRM (MML Conc Obs)	GRM (MML Conc Obs)	GRM (MML Conc Obs)
	40%	PCM (JML Conc True)	Odds-transform (Chained SD)	GRM (MML Conc Obs)	GRM (MML Conc Obs)	GRM (MML Conc Obs)
Random	18%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	25%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	33%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	40%	Odds-transform (Tucker SD)	Transformed Linear (Tucker)	GRM (MML Conc Obs)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)

Table 11. The best-performing classical equating methods under each simulation scenario (i.e., those with the smallest mean weighted RMSE once Rasch/IRT methods are excluded).

Group difference	Form overlap	Sample size				
		50	100	500	1000	2000
Maximum	18%	Symmetric circle-arc (Nominal)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)
	25%	Symmetric circle-arc (Nominal)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)
	33%	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)
	40%	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)	Symmetric circle-arc (Levine)
Medium	18%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	Transformed Linear (Chained)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)
	25%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	Odds-transform (Levine SD)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)
	33%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	Chained Equi (Kernel smoothed)	FE Equi (Modified WB Unsmoothed)	FE Equi (Modified WB Unsmoothed)
	40%	Odds-transform (Chained SD)	Odds-transform (Chained SD)	Chained Equi (Kernel smoothed)	Chained Equi (Kernel smoothed)	FE Equi (Modified WB Unsmoothed)
Random	18%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	25%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	33%	Odds-transform (Tucker SD)	Odds-transform (Tucker SD)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)
	40%	Odds-transform (Tucker SD)	Transformed Linear (Tucker)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)	FE Equi (Kernel smoothed)

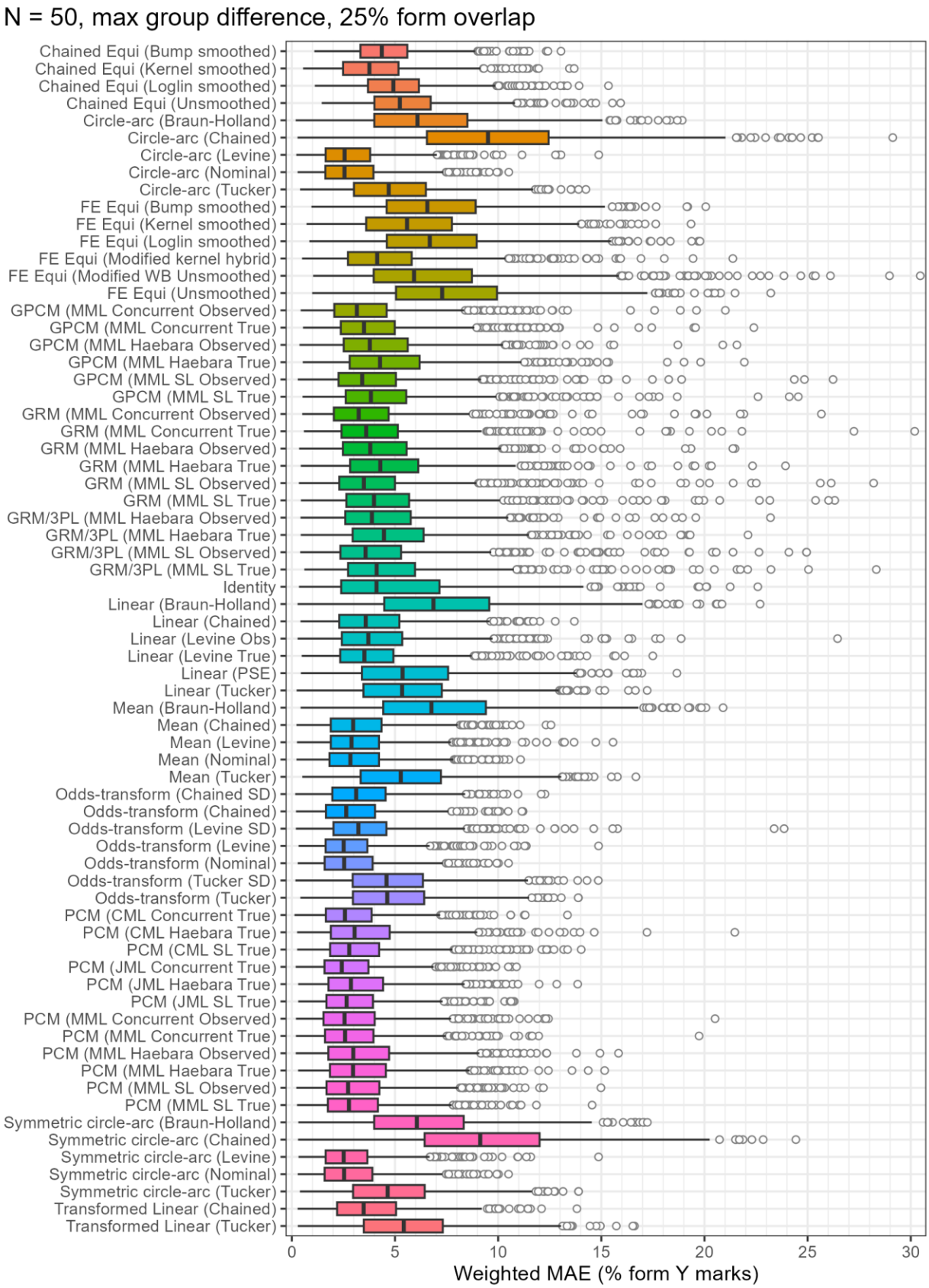

Figure 38. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 50, 25% form overlap and a large difference between groups. Note that outliers with values >30% are not shown.

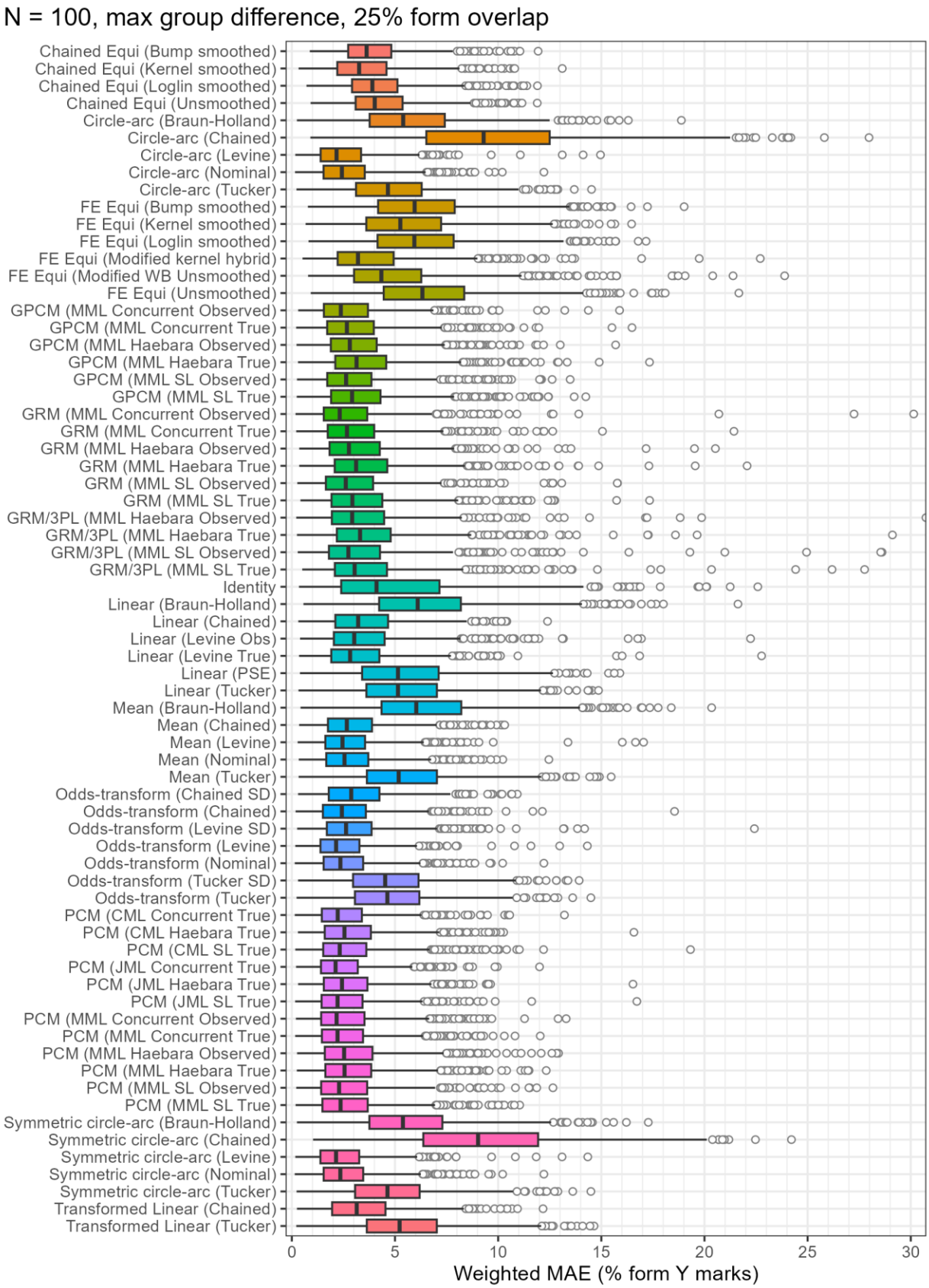

Figure 39. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 100, 25% form overlap and a large difference between groups. Note that outliers with values >30% are not shown.

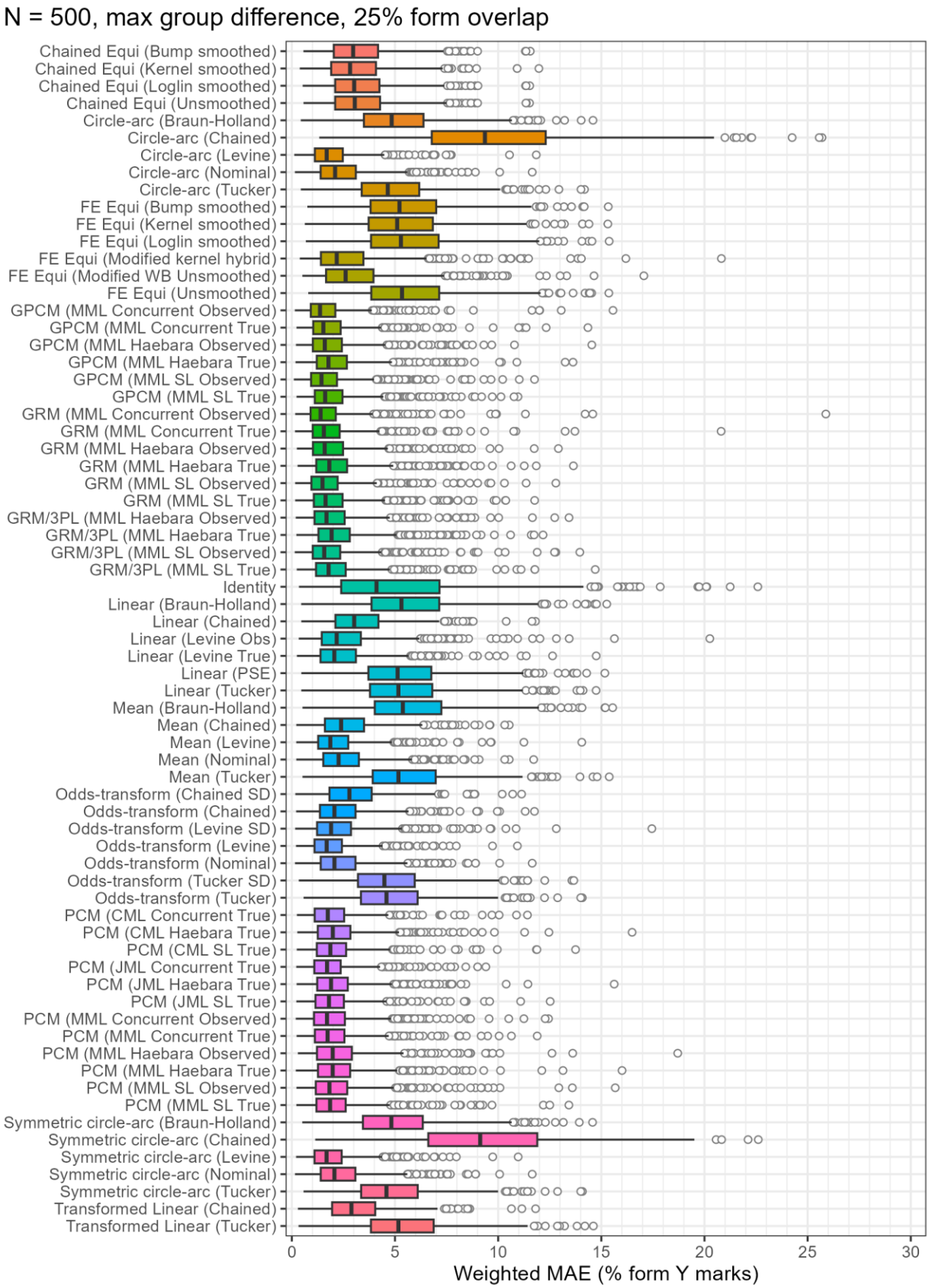

Figure 40. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 500, 25% form overlap and a large difference between groups. Note that outliers with values >30% are not shown.

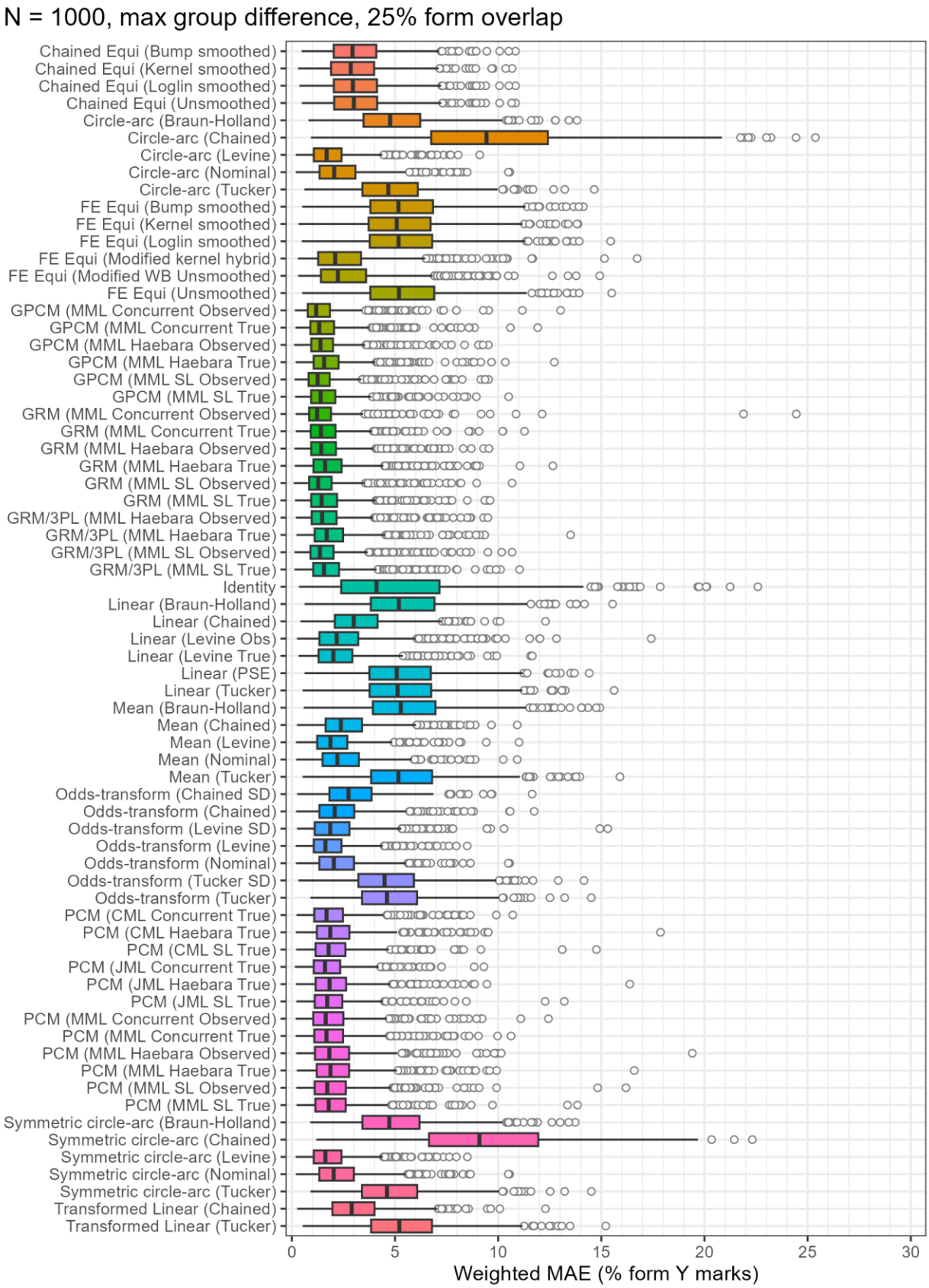

Figure 41. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 1,000, 25% form overlap and a large difference between groups. Note that outliers with values >30% are not shown.

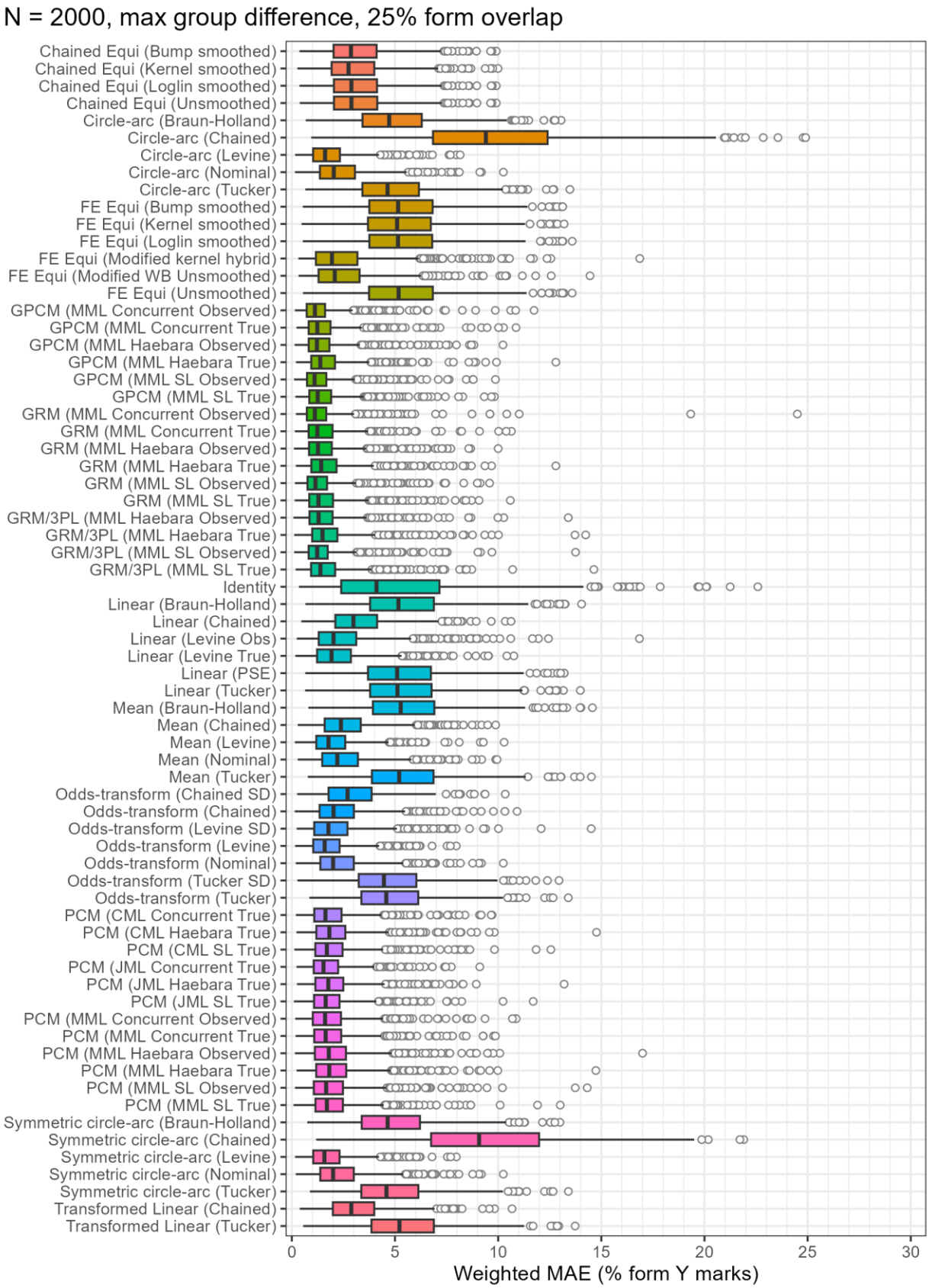

Figure 42. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 2,000, 25% form overlap and a large difference between groups. Note that outliers with values >30% are not shown.

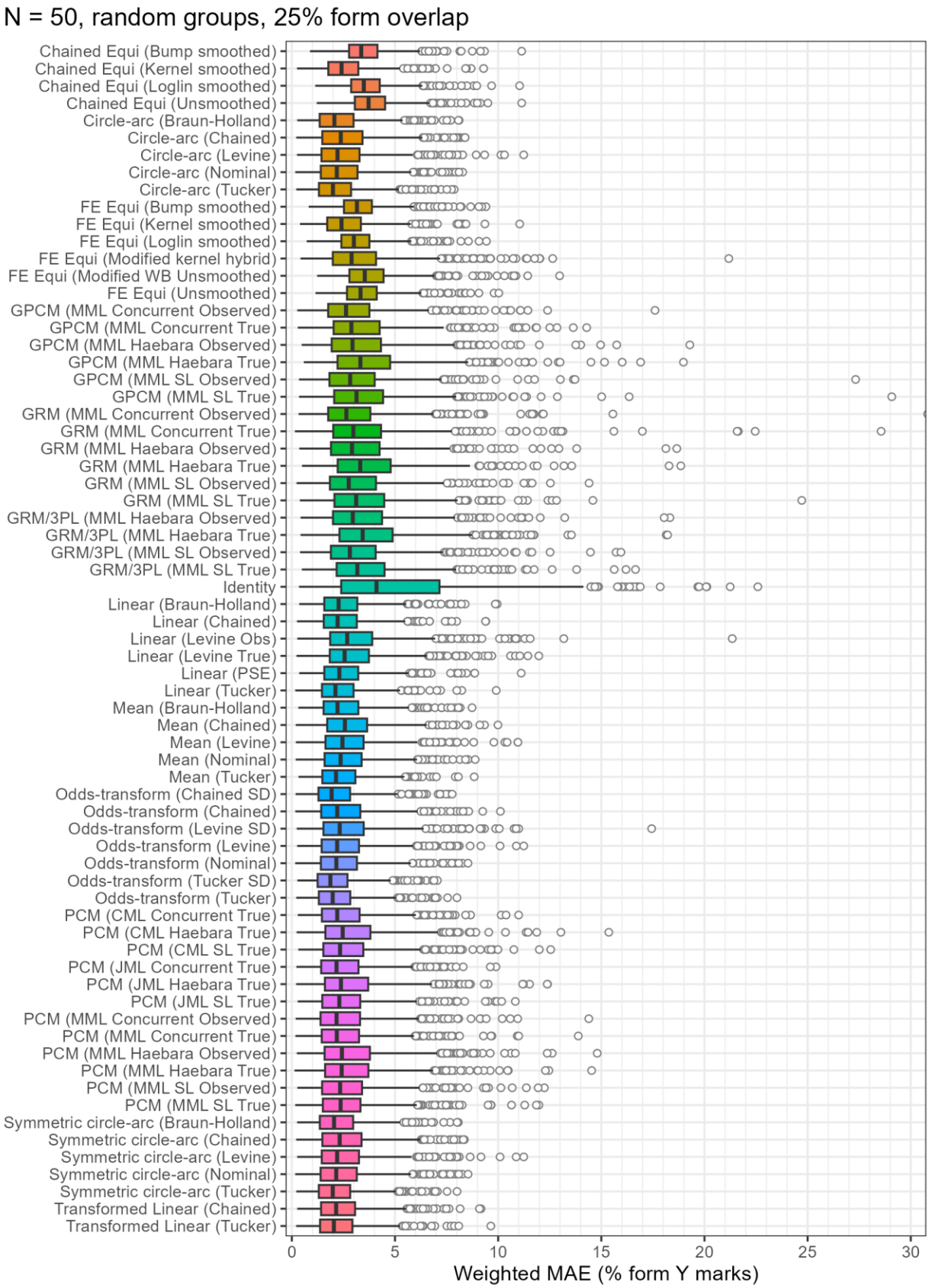

Figure 43. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 50, 25% form overlap and random groups. Note that outliers with values >30% are not shown.

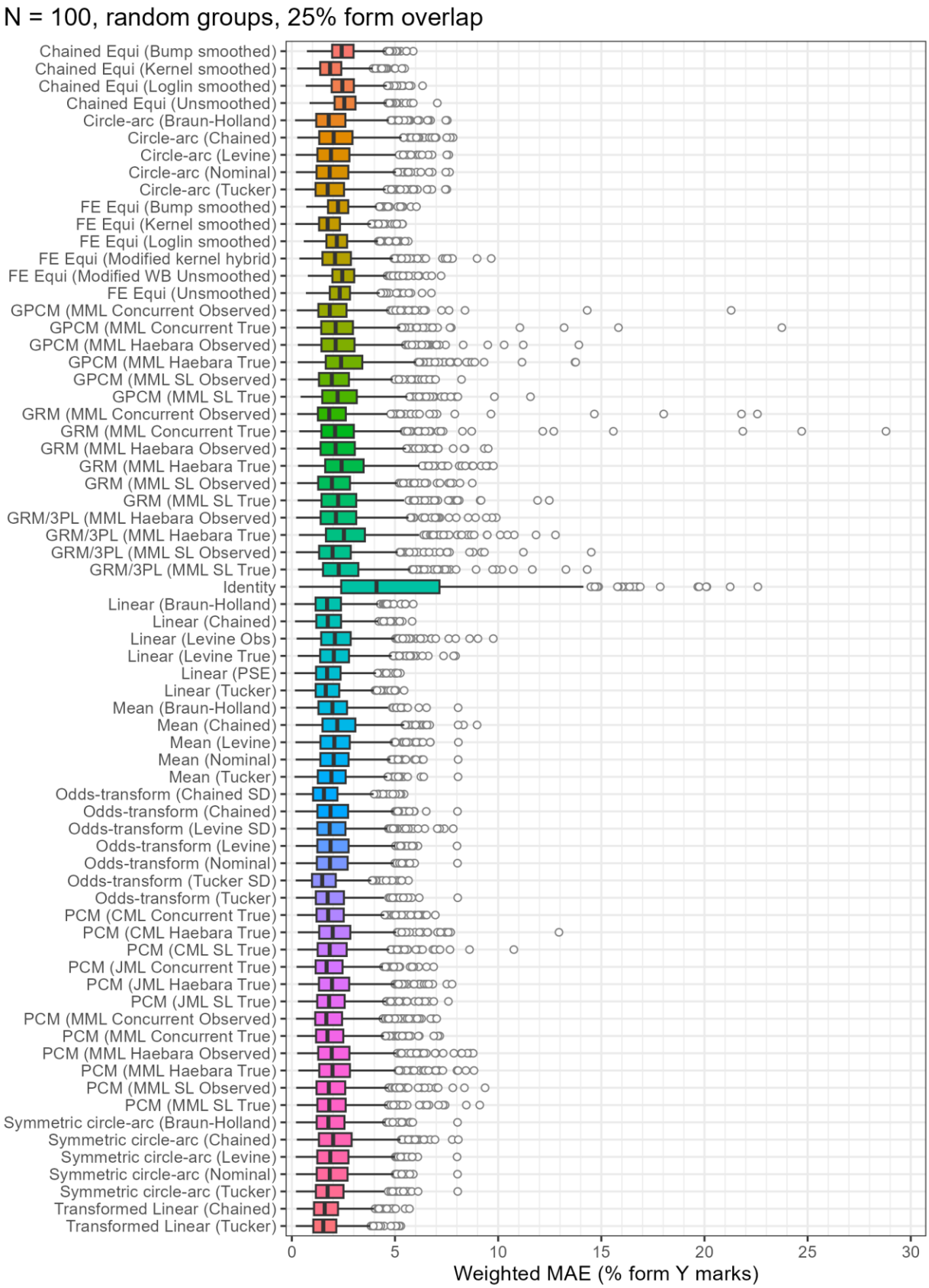

Figure 44. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 100, 25% form overlap and random groups. Note that outliers with values >30% are not shown.

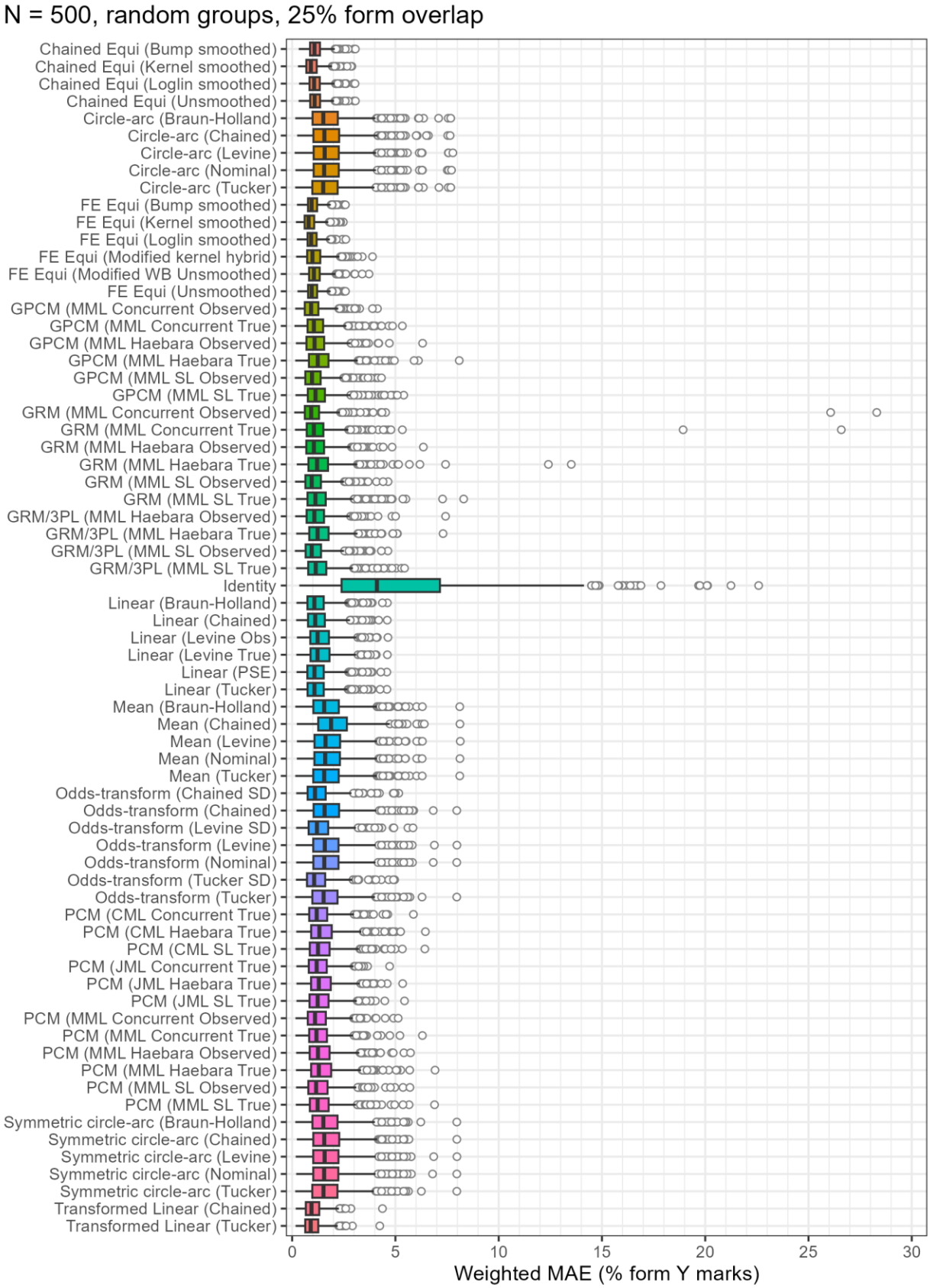

Figure 45. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 500, 25% form overlap and random groups. Note that outliers with values >30% are not shown.

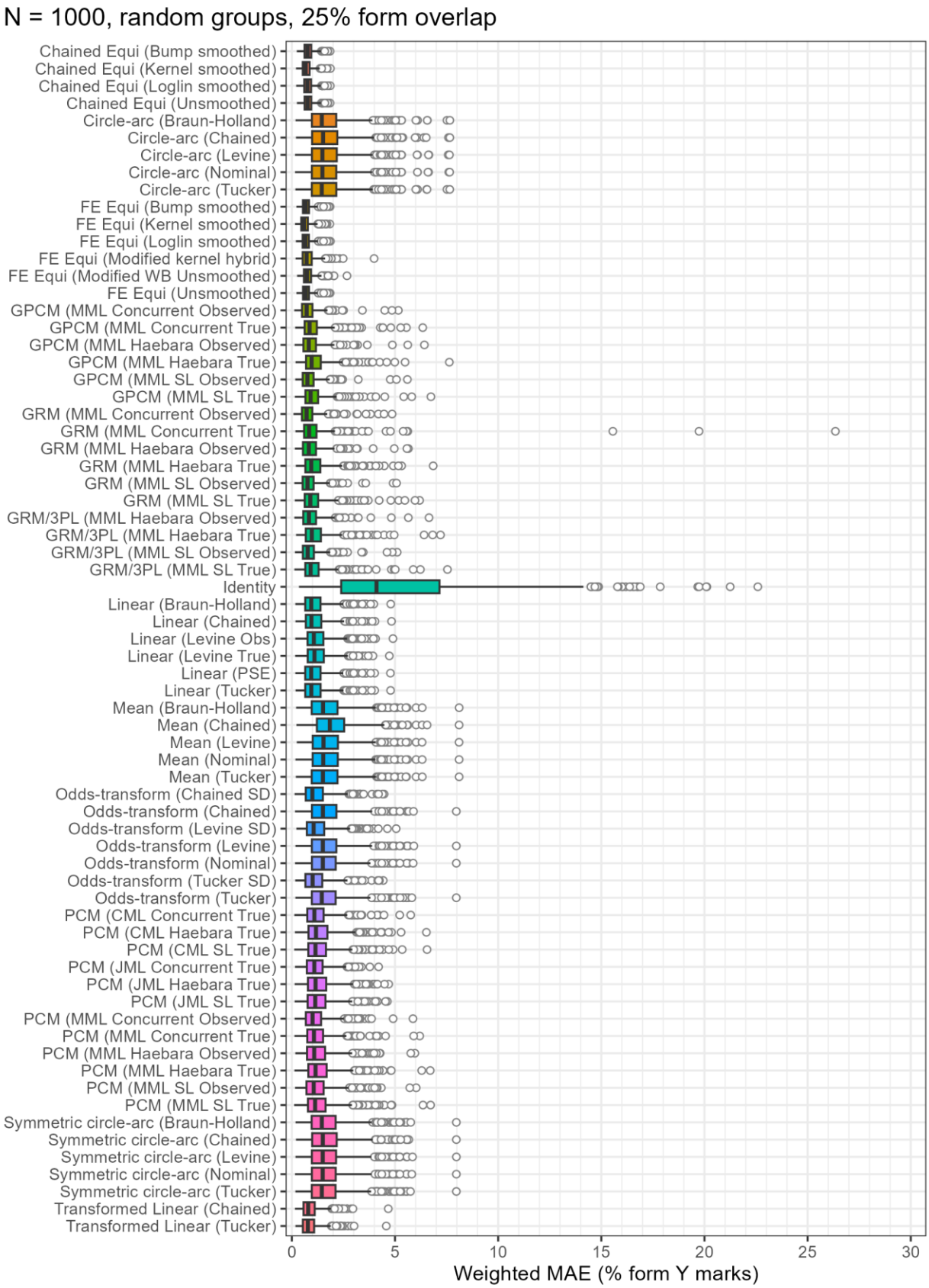

Figure 46. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 1,000, 25% form overlap and random groups. Note that outliers with values >30% are not shown.

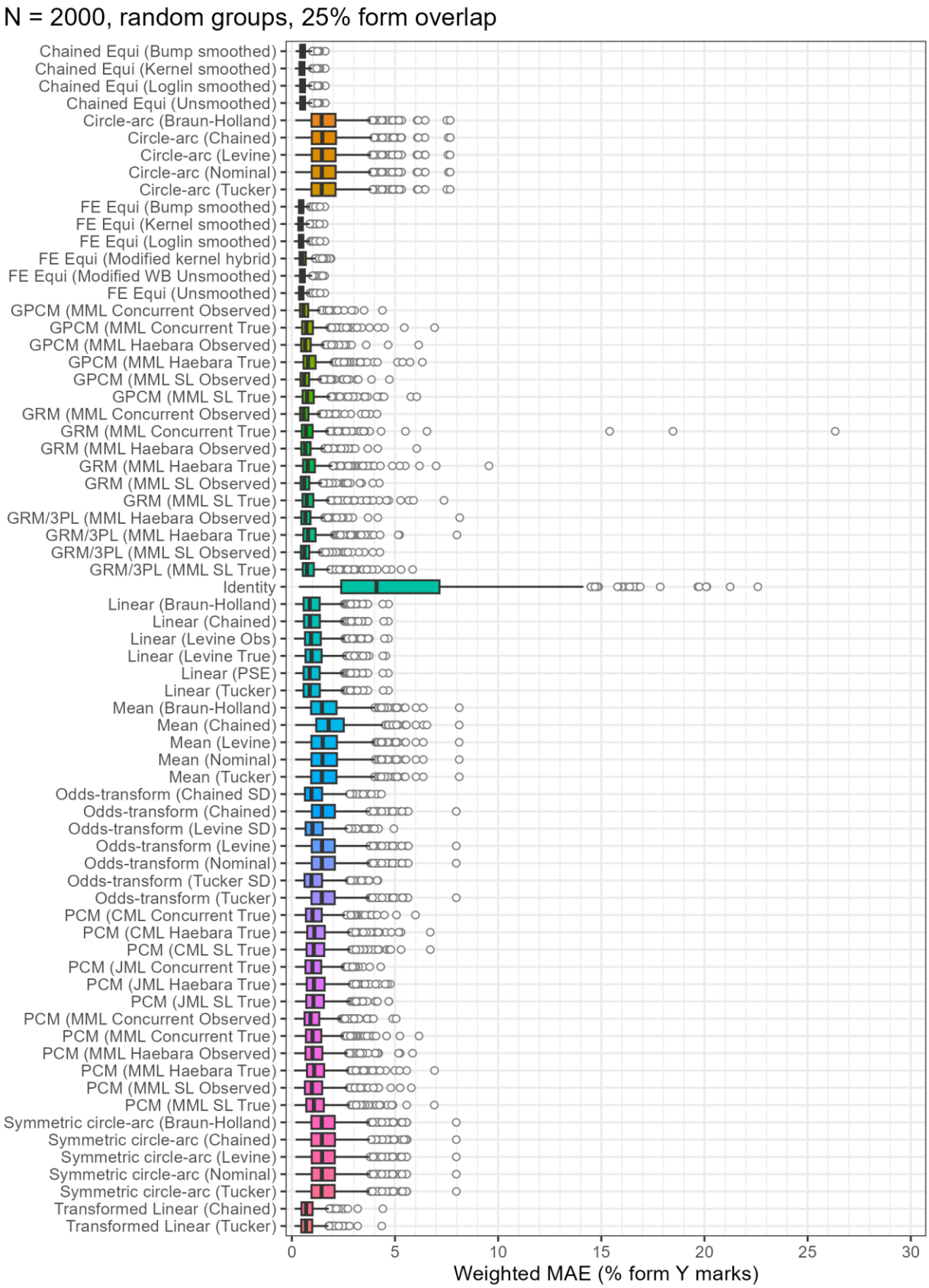

Figure 47. Box plots of weighted MAE values (expressed as % of Form Y marks) for all equating methods tested, for a sample size of 2,000, 25% form overlap and random groups. Note that outliers with values >30% are not shown.