

UC Merced

UC Merced Previously Published Works

Title

The NEAT Equating Via Chaining Random Forests in the Context of Small Sample Sizes: A Machine-Learning Method.

Permalink

<https://escholarship.org/uc/item/49s6x2nn>

Journal

Educational and Psychological Measurement, 83(5)

Authors

Jiang, Zhehan

Han, Yuting

Xu, Lingling

et al.

Publication Date

2023-10-01

DOI

10.1177/00131644221120899

Peer reviewed

The NEAT Equating Via Chaining Random Forests in the Context of Small Sample Sizes: A Machine-Learning Method

Educational and Psychological
Measurement

2023, Vol. 83(5) 984–1006

© The Author(s) 2022




Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/00131644221120899

journals.sagepub.com/home/epm



Zhehan Jiang¹, Yuting Han¹, Lingling Xu¹ , Dexin Shi² ,
Ren Liu³ , Jinying Ouyang¹ and Fen Cai¹

Abstract

The part of responses that is absent in the nonequivalent groups with anchor test (NEAT) design can be managed to a planned missing scenario. In the context of small sample sizes, we present a machine learning (ML)-based imputation technique called chaining random forests (CRF) to perform equating tasks within the NEAT design. Specifically, seven CRF-based imputation equating methods are proposed based on different data augmentation methods. The equating performance of the proposed methods is examined through a simulation study. Five factors are considered: (a) test length (20, 30, 40, 50), (b) sample size per test form (50 versus 100), (c) ratio of common/anchor items (0.2 versus 0.3), and (d) equivalent versus nonequivalent groups taking the two forms (no mean difference versus a mean difference of 0.5), and (e) three different types of anchors (random, easy, and hard), resulting in 96 conditions. In addition, five traditional equating methods, (1) Tucker method; (2) Levine observed score method; (3) equipercenile equating method; (4) circle-arc method; and (5) concurrent calibration based on Rasch model, were also considered, plus seven CRF-based imputation equating methods for a total of 12 methods in this study. The findings suggest that benefiting from the advantages of ML techniques, CRF-based methods that incorporate the equating result of the Tucker method, such

¹Peking University Health Science Center, Beijing, China

²University of South Carolina, Columbia, USA

³University of California, Merced, USA

Corresponding Author:

Yuting Han, Institute of Medical Education, Peking University Health Science Center, 38 Xueyuan Road, Beijing 100191, China.

Email: hanyuting@bjmu.edu.cn

as IMP_total_Tucker, IMP_pair_Tucker, and IMP_Tucker_cirlce methods, can yield more robust and trustable estimates for the “missingness” in an equating task and therefore result in more accurate equated scores than other counterparts in short-length tests with small samples.

Keywords

small samples, equating, chaining random forests, machine learning–based imputation techniques

Introduction

In educational settings, producing interchangeable scores on different test forms (i.e., equating) is essential to make the assessment fair and comparable when examining unidentical items/questions (Kolen & Brennan, 2004). A majority of researchers and practitioners perform equating through the nonequivalent groups with an anchor test (NEAT) design, which adjusts items’ properties to estimate what an examinee would have performed if this examinee was administered items that were, in fact, never administered (Maris et al., 2010). To illustrate without losing generalizability, a typical NEAT design with two forms made of three batches of items is given here: one batch for the base/reference form only, the second batch for the target form only, and the third batch shared between the forms (i.e., the anchor set). Traditionally, statistical techniques for equating are about transformations of both modeling parameters and item responses, including the ones based on equipercentile equating, linear equating methods, item response theory (IRT) observed score and true score equating, van der Linden local equating, Levine nonlinear method, Kernel equating (KE), and others (see Kolen & Brennan, 2004 for details). Furthermore, post-stratification (PSE), Levine observed score linear, and chained equating (CE) methods are typically used in KE when a NEAT design is present (von Davier et al., 2004). In addition to treating equating as the transformation, it can also be handled as a missing data problem.

Following a popular definition of missing data in the statistics literature, the part of responses absent in a NEAT design can be regarded as missing at random, known as missing at random (MAR) mechanism (Little & Rubin, 2002). Accordingly, values underlying the missing areas depend on the design part, of which the responses are observable. On the other hand, Sinharay and Holland (2010) claimed that since the missingness in the NEAT design is deliberately planned, and therefore theoretically, it is likely to be missing completely at random (MCAR) instead of from a theoretical perspective. We believe this assumption applies in most cases, except possibly affected by testing time. Methodologically speaking, techniques for handling missing data problems can be applied to both MAR and MCAR settings. Previous studies have treated the NEAT design as an incomplete-data issue (Liou & Cheng, 1995; Liou et al., 2001), involving imputation methods designated for MAR problem,

including a kernel estimator, a log-linear model-based estimator, and an iterative moment estimator.

Conventionally, imputation techniques can be either model-based or model-free. Readers interested in comprehensive imputation approaches can see Little and Rubin (2019) and Enders (2010) for details. To illustrate the model-based one related to equating tasks, Holland and Thayer (2000) proposed an algorithm based on “expectation-maximization” (EM), where the aforementioned log-linear model was deployed to produce values for the missing part (i.e., the equating target), and Moses and colleagues (2011) found that the approach was fairly reliable in many NEAT conditions. On the other hand, the model free-based imputations (i.e., K-nearest neighbors, fuzzy K-means (i.e., an extension of K-means that does not simply predict targets to a definitive class but provide class-probability estimates like a mixture model; Bezdek, 1981; Equihua, 1990), singular value decomposition, principal component analysis, and others) seem to be less favored in this kind of study of multiple imputations by chained equations (MICE). These model-free approaches are now labeled machine learning (ML)-based imputation techniques in the contemporary world (Lakshminarayan et al., 1996; Lin & Tsai, 2020), emphasizing predictive accuracy rather than interpretability.

Similar to other ML-based approaches, the advantages such as needing minimal assumptions about the data-generating systems, being compatible with complex variable patterns, subsuming various input formats, as well as producing more trustable predictions make ML-based imputation techniques a popular choice in both research and practice (Athey, 2018; Ij, 2018), especially in the conditions where simple linear associations between the missing and the observed data do not exist (Hong et al., 2020). These properties make ML-based imputation techniques promising for equating tasks. As listed in Figure 1, visual comparison bridges the essence of equating tasks and imputing inquiries. Group X takes test form #1, while group Y takes test form #2, and common items designed to be the same in both tests are called anchor items. True responses from Group X on test form #2 and Group Y on test form #1 are missing except for anchor items. Imputation techniques used to deal with missing data can be used to obtain equating scores for individuals on unanswered tests.

The primary inquiry of equating is about yielding more accurate estimates for hypothetical scores obtained from the base form that an examinee never actually takes. This inquiry matches the ML advantages mentioned above well and, therefore, inspires the possibility of applying the techniques to situations where traditional equating approaches commonly used in testing organizations fail to deliver reliable results. Unsurprisingly, small sample equating is one of the situations; it has received more attention in the literature nowadays. The literature review found that linear equating methods have been suggested for use with small samples (Kolen & Brennan, 2004; Skaggs, 2005). In addition, several new methods for small-sample equating have been proposed, including circle-arc equating (Livingston & Kim, 2009), synthetic equating (Kim et al., 2008), nominal weights mean equating (Babcock et al., 2012), and so on.

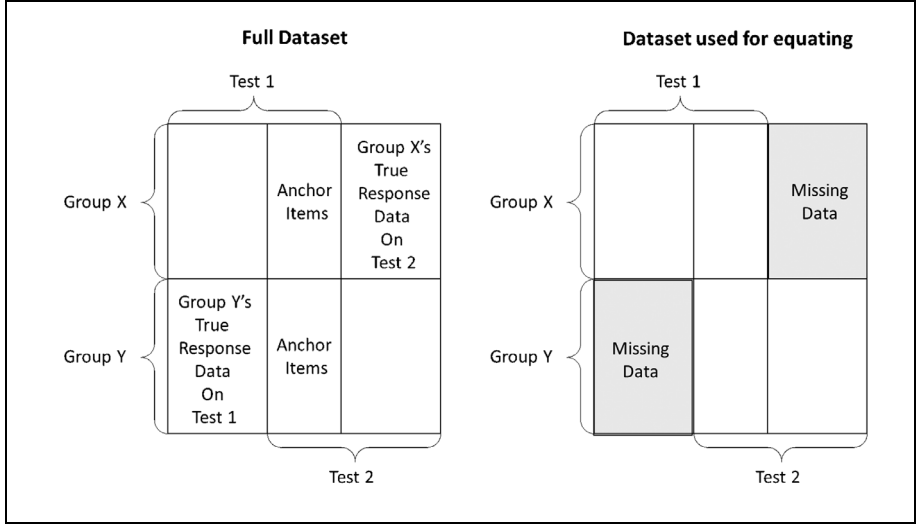


Figure 1. The Bridge Between Equating Task and Imputation.

Recent studies addressing small sample equating have primarily focused on evaluating the performance of existing methods. For instance, circle-arc equating and nominal weights mean equating yielded less-biased estimates in small samples compared to applications within standard settings for each administration (Dwyer, 2016). A recent study by Babcock and Hodge (2020) showed that Rasch-based approaches could produce acceptable results in the context of small sample exams, especially for non-Bayesian ones.

In this study, we propose using an ML-based imputation technique called chaining random forests (CRF) to perform equating tasks within a NEAT design, given a scenario of small sample sizes, defined as a low volume of both examinees and items. The equating performance of the proposed methods was also compared with other equating methods likely to be used in small-sample situations through a simulation study. We hypothesized that by benefiting from ML techniques' advantages, CRF would yield more robust estimates for the "missingness" in an equating task and therefore result in more reliable equated scores than other counterparts.

Method

Initially introduced by Stekhoven and Bühlmann (2012), CRF is an iterative imputation technique devising Breiman's random forest algorithm (Breiman, 2001). As CRF's name suggests, the major components are random forests, trained on the observed values to predict the missing ones. The advantage of this method is that it considers complex interactions and non-linear relations among variables. Studies have shown that CRF and MICE can be equivalent in many situations, where the

former is not only a better fit for mixed-type data (Penone et al., 2014; Yadav & Roychoudhury, 2018) but also consumes less computational power when a similar task is present (Wong et al., 2021). Substantial evidence has published to support this study, for instance, Shah and colleagues (2014) compared random forest to multiple imputation by chained equations (MICE) and showed that random forest parameters were less biased.

Consider that in a data set where an arbitrary variable \mathbf{y}_s contains missingness at entries $\mathbf{i}_{mis}^{(s)} \in \{1, \dots, n\}$, where \mathbf{y}_s can be viewed as the response vector for all n subjects on item s . $\mathbf{i}_{obs}^{(s)}$ is the complement of $\mathbf{i}_{mis}^{(s)}$ at all entities, representing the index of individuals with no missing on \mathbf{y}_s . The data set can be classified into four subsets: (1) the observed values of variable \mathbf{y}_s denoted by $\mathbf{y}_{obs}^{(s)}$, (2) the missing values of variable \mathbf{y}_s denoted by $\mathbf{y}_{mis}^{(s)}$, (3) the variables other than \mathbf{y}_s with observations $\mathbf{i}_{obs}^{(s)}$ denoted by $\mathbf{X}_{obs}^{(s)}$, and (4) the variables other than \mathbf{y}_s with observations $\mathbf{i}_{mis}^{(s)}$ denoted by $\mathbf{X}_{mis}^{(s)}$. It should be noted that $\mathbf{X}_{obs}^{(s)}$ does not imply the corresponding values are completely observed, as the index $\mathbf{i}_{obs}^{(s)}$ corresponds to the observed values of the variable \mathbf{y}_s . Similarly, $\mathbf{X}_{mis}^{(s)}$ is not completely missing neither. The algorithm starts by having initial values for all missing areas. Then, the variables \mathbf{y}_s are sorted for $s = 1, \dots, p$ according to their missing proportions. For every \mathbf{y}_s , the imputation is achieved by using a random forest (RF) with output $\mathbf{y}_{obs}^{(s)}$ and input $\mathbf{X}_{obs}^{(s)}$; the trained RF is then used to predict $\mathbf{y}_{mis}^{(s)}$ from $\mathbf{X}_{mis}^{(s)}$. This process is iterated until a stopping criterion is met (see Stekhoven & Bühlmann, 2012 for algorithm details).

To eventually deploy CRF in equating tasks, we defined six ways of augmentations for the imputations and used the original data set to impute all missing values as a baseline method (IMP); the corresponding implementations are listed in Figure 2. Specifically, the first augmentation incorporated each student's total score on anchor items as a new column into the data (IMP_total). The second one augments the data by adding the sum scores of each item pair nested within the anchor test (IMP_pair). The latter four data augmentation methods were constructed by exploiting benefits from well-known equating methods (e.g., the Tucker method and the circle-arc method); these augmentation methods can be divided into two steps: (1) the equating procedure (the Tucker method or the circle-arc method) is first implemented to calculate the equating scores of the target group (group Y) on the reference test (test form #1), and (2) the equating scores and the total scores of the reference group (group X) on the reference test (test form #1) are combined to form a new variable to augment the original data set. The method using both the total anchor test score and the equating results of the Tucker method is named IMP_toatl_Tucker method, while the one using both the sum scores of each item pair in the anchor test and the equating results of the Tucker method is called IMP_pair_Tucker method. To compare the outcomes using different equating methods, the method called IMP_toatl_circle (i.e., using both the total anchor test score and information from the circle-arc method) was used for comparison. Finally, total scores of the anchor test, sum scores of each item pair in

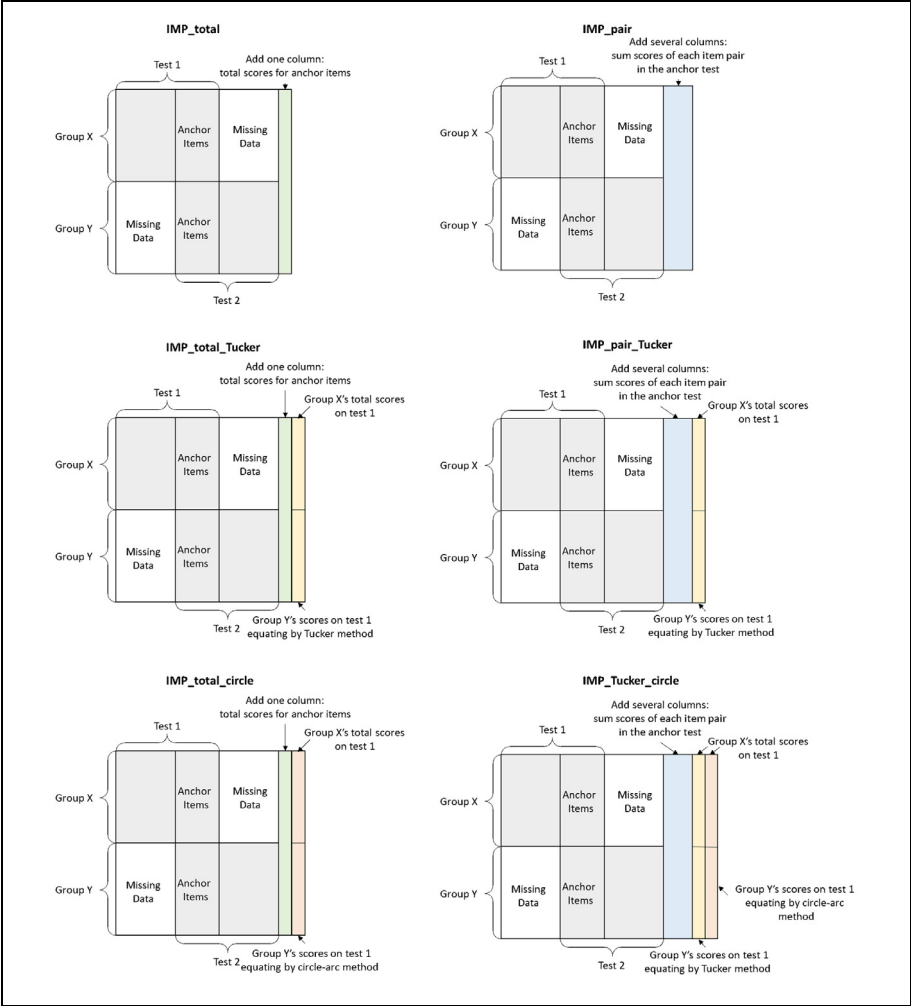


Figure 2. Six Methods of Augmentation for Imputations.

the anchor test, and the equating results from both the Tucker and the circle-arc methods were added to the data simultaneously to form IMP_Tucker_circle method to investigate if the equating performance could be further improved by using more information.

Simulation Study

When comparing different equating methods in simulation studies of this kind, common factors include the sample size (Arai & Mayekawa, 2011; Hanson & Béguin,

2002; Kang & Petersen, 2012; Kim & Cohen, 1998; Sinharay & Holland, 2007), the number or proportion of anchor items (Arai & Mayekawa, 2011; Hanson & Béguin, 2002; Kang & Petersen, 2012; Kim & Cohen, 1998; Sinharay & Holland, 2007; T. Wang et al., 2008), the ability distribution of both the target group and reference group (Hanson & Béguin, 2002; Kang & Petersen, 2012; Kim & Cohen, 1998; Sinharay & Holland, 2007; T. Wang et al., 2008), the difficulty distribution of anchor items (Hanson & Béguin, 2002; Kang & Petersen, 2012; Sinharay & Holland, 2007), and the test length (Sinharay & Holland, 2007; T. Wang et al., 2008).

Summarizing the aforementioned designs to accommodate the small sample context (e.g., a classroom setting; Perry & Dickens, 1987; Stewart & Gibson, 2010), this study considered the following factors:

1. Test length. Four levels of test length were considered: 20, 30, 40, and 50.
2. Sample size. The sample sizes for X and Y were equally set with two levels: 50 and 100.
3. Proportion of anchor items: 0.2 and 0.3.
4. Two (latent) ability distributions for the target group Y: $N(0,1)$ and $N(0.5,1)$.
5. Difficulty distribution of anchor items: random, easy, and hard. When the type of anchor items was random, the anchor items were randomly selected from test form #1. Otherwise, the difficulty distribution of anchor items was biased from test form #1. When the type of anchor items was easy, the anchor items were randomly selected from half of the items with lower difficulty values from test form #1. Conversely, when the type of anchor items was difficult, the anchor items were randomly selected from half of the items with higher difficulty values in test form #1.

There were 96 conditions in total (i.e., $4 \times 2 \times 2 \times 2 \times 3$). The three-parameter logistic (3PL) IRT model (Birnbaum, 1968) was adopted for data generation in multiple conditions via the NEAT design, assuming that group X took test form #1 and group Y took test form #2. While the specific procedures can be found in Online Appendix, the simulation and analyses involved the following steps:

Step 1: The discrimination parameters, difficulty parameters, and guessing parameters of both tests (test form #2 did not include anchor items at this step) were randomly generated from $N(0.8, 0.2)$, $N(0, 1)$, and $Unif(0, 0.25)$.

Step 2: Ability values for group X were randomly generated from the standard normal distribution $N(0,1)$. The ability values for group Y were generated according to its factor levels. The full data set was generated using the IRT model based on the item parameters and ability values.

Step 3: Sort the items in test form #1 according to their difficulty values. A pre-defined number (according to the simulation condition) of anchor items was randomly selected from test form #1 in alignment with their difficulty levels.

Table 1. Descriptive Statistics for the Difficulty Parameter of the Anchor Items.

Anchor type	Statistic	M	Minimum	Maximum
Easy	M (b)	-0.566	-1.161	0.113
	SD (b)	0.663	-0.118	1.199
Random	M (b)	0.109	-0.931	1.104
	SD (b)	0.853	0.170	1.740
Hard	M (b)	0.798	0.490	1.294
	SD (b)	0.381	0.094	0.624

The responses of group X on test form #2 and group Y on test form #1 were treated as true observed data and set as missing data, as shown in Figure 1.

Step 4: Given group X’s responses on test form #1 and group Y’s responses on test form #2, different equating methods were used to compute equivalent scores converted from test form #2 to test form #1.

Step 5: Steps 2 to 4 were repeated 100 times. Two measures were used according to the literature (i.e., Wolkowitz & Wright, 2019; Zeng, 1993)—the average absolute bias (BIAS) and root mean square difference (RMSD):

$$BIAS = \frac{\sum_p^N |X_p \text{ equated} - X_p \text{ observed}|}{N} \tag{1}$$

$$RMSD = \sqrt{\frac{\sum_p^N (X_p \text{ equated} - X_p \text{ observed})^2}{N}} \tag{2}$$

where N is the total number of examinees taking test form #1, $X_p \text{ equated}$ is a person’s equated score converted from test form #2 onto test form #1, and $X_p \text{ observed}$ is a person’s observed score on test form #1 in the simulated data. Based on the repeated samples, the measures were calculated by averaging over 100 repetitions.

Step 6: Steps 1 to 5 were repeated for each simulation condition, where the indexes were recorded for further comparisons.

The descriptive statistics of the mean and standard deviation of the difficulty for the anchor items under different anchor types are shown in Table 1.

To compare the proposed method with equating methods commonly used in large-scale testing organizations, especially those performed better with small samples, linear equating methods (Tucker method and Levine observed score method), equipercentile equating method, circle-arc method and concurrent calibration (CC)

Table 2. Average Equating Errors From Different Equating Methods.

Equating method	RMSD	BIAS
IMP	4.021	3.227
IMP_total	3.971	3.185
IMP_pair	3.917	3.149
IMP_total_Tucker	3.235	2.609
IMP_pair_Tucker	3.348	2.699
IMP_total_circle	3.235	2.609
IMP_Tucker_circle	3.236	2.610
Tucker	3.314	2.643
Levine	3.455	2.748
equipercntile	3.398	2.708
circle-arc	3.367	2.686
Rasch	3.398	2.710

Note. The smallest values among the equating methods are boldfaced. RMSD = root mean square difference.

(Hanson & Béguin, 2002; Hu et al., 2008) based on Rasch model with true score equating (Kolen & Brennan, 2004) were selected to serve as references. R software was used (R Core Team, 2022), while the packages “equate” (Albano, 2016), “equateIRT” (Battauz, 2015), “SNSequate” (González, 2014), and “missRanger” (Mayer & Mayer, 2022) were implemented to execute the reference methods and CRF imputation, respectively. All the package settings were left default. The R script for data generation, imputation/equating, and result gathering was documented in the Appendix.

Result

Aggregated results are presented across all conditions in Table 2. The smallest RMSD and BIAS values among the equating methods are boldfaced. The three imputation methods using only raw data or its integrated information (IMP, IMP_total and IMP_pair method) did not perform as well as the traditional equating methods: those that used data augmentations (IMP_total and IMP_pair) yielded smaller RMSD values than the one used the original data set only (i.e., IMP). Using the sum of item pairs to augment the data (IMP_pair) was slightly better than using the total scores of anchor items (IMP_total). Meanwhile, adding information from other equating methods significantly improved the performance of the imputation methods, with a significant decrease in both averaged RMSD and BIAS. IMP_total_Tucker and IMP_total_circle outperformed the other methods as they had the lowest RMSD and BIAS values. IMP_pair_Tucker method, which also uses Tucker method information, was inferior to IMP_total_Tucker method. In addition, the equating accuracy could not be further enhanced when multiple sources were used together (i.e., the total scores of

anchor items, the sum of anchor item pairs, and the equating results of both Tucker and circle-arc methods) to augment the data (IMP_total_circle), and its RMSD and BIAS were close to those of IMP_total_Tucker method.

The Tucker method performed best among the traditional equating methods because it had the lowest RMSD and BIAS values, followed by the circle-arc method. The equipercentile equating and the Rasch-based methods both performed poorly because they produced the largest RMSD value, and the Levine method generated the largest BIAS among the traditional methods.

The results for the average equating errors of the 12 equating methods across different test conditions are presented in Tables 3 to 10 and discussed in the following paragraphs.

Test Length and Sample Size

A comparison of Tables 3 to 6 and Tables 7 to 10 indicates that all the equating methods followed a similar pattern, where they tended to produce larger RMSD and BIAS values as the number of items on a test form increased. When the test length was 20 and 30, imputation methods incorporating results from other methods (i.e., IMP_total_Tucker, IMP_pair_Tucker, IMP_total_circle and IMP_Tucker_circle) outperformed any other single (non-incorporated) equating methods. Among them, IMP_Tucker_circle method, which uses more information, was more accurate than its counterparts in most cases. When the test length was 40, the Tucker method started to show a slight advantage on a small set of conditions, and when the test length reached 50, the advantage became more apparent; however, imputation methods, such as IMP_total_Tucker and IMP_total_circle methods, still produced the highest equating accuracy in some cases. In addition, the advantage of IMP_Tucker_circle method over IMP_total_Tucker and IMP_total_circle methods faded apart.

The equating accuracy for traditional equating methods tends to improve as the sample size increases. When the sample size was 100, they produced smaller RMSD and BIAS values in most cases than when the sample size was 50. On the contrary, for CRF-based imputation methods, the effect of sample sizes was inconsistent: they yielded larger RMSD or BIAS values when the sample size became larger in several conditions, except for IMP_pair_Tucker method, of which the equating accuracy improved as sample size increased. Although the performance of IMP_pair_Tucker method was not as good as IMP_total_Tucker method on average, the RMSD and BIAS values of IMP_pair_Tucker method were lower than those of IMP_total_Tucker method in cases of 100 examinees, especially when the test length was 20: the RMSD and BIAS values of IMP_total_Tucker method were always lower than those of IMP_pair_Tucker method and even the smallest among all equating methods in most cases.

Table 3. The Overall RMSD for Different Equating Methods (Number of Items = 20).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches										Linear method		
				IMP	IMP total	IMP pair	IMP Tucker	IMP total Tucker	IMP pair Tucker	IMP circle	IMP Tucker circle	Tucker	Levine	Equipper-centile	Circle-arc	Rasch
0.2	50	rand	0	2.947	2.967	2.985	2.409	2.422	2.402	2.401	2.663	2.979	2.708	2.661	2.709	
			0.5	2.976	2.989	2.972	2.417	2.416	2.408	2.404	2.637	2.953	2.699	2.643	2.685	
		hard	0	3.019	3.031	3.030	2.453	2.459	2.440	2.430	2.667	2.822	2.716	2.668	2.711	
			0.5	3.062	3.067	3.050	2.394	2.417	2.385	2.388	2.620	2.863	2.656	2.610	2.661	
		easy	0	2.919	2.936	2.940	2.416	2.413	2.423	2.411	2.688	2.919	2.705	2.706	2.764	
			0.5	2.973	2.963	2.964	2.403	2.403	2.401	2.404	2.641	2.985	2.673	2.676	2.731	
	100	rand	0	2.995	3.004	2.994	2.405	2.386	2.423	2.404	2.608	2.740	2.618	2.626	2.664	
			0.5	3.001	3.009	2.997	2.401	2.377	2.419	2.401	2.620	2.724	2.636	2.625	2.658	
		hard	0	3.048	3.060	3.054	2.433	2.400	2.427	2.408	2.636	2.807	2.654	2.603	2.653	
			0.5	3.017	3.011	3.008	2.431	2.389	2.438	2.400	2.566	2.662	2.588	2.561	2.601	
		easy	0	2.895	2.908	2.900	2.404	2.368	2.406	2.406	2.625	2.712	2.636	2.662	2.729	
			0.5	2.986	2.992	2.983	2.423	2.398	2.423	2.410	2.630	2.765	2.635	2.654	2.719	
0.3	50	rand	0	2.586	2.579	2.585	2.210	2.237	2.222	2.196	2.457	2.608	2.503	2.488	2.512	
			0.5	2.592	2.596	2.581	2.200	2.226	2.192	2.176	2.464	2.593	2.498	2.473	2.506	
	hard	0	2.592	2.559	2.577	2.238	2.230	2.246	2.206	2.431	2.549	2.458	2.454	2.467		
		0.5	2.588	2.586	2.597	2.260	2.229	2.260	2.209	2.404	2.571	2.441	2.418	2.412		
	easy	0	2.489	2.491	2.515	2.207	2.208	2.199	2.175	2.442	2.493	2.458	2.508	2.560		
		0.5	2.533	2.536	2.532	2.213	2.204	2.214	2.171	2.441	2.498	2.465	2.516	2.569		
100	rand	0	2.531	2.552	2.567	2.242	2.186	2.236	2.184	2.409	2.484	2.428	2.441	2.462		
		0.5	2.537	2.554	2.564	2.269	2.189	2.276	2.215	2.400	2.437	2.416	2.440	2.454		
	hard	0	2.626	2.633	2.662	2.336	2.254	2.333	2.273	2.432	2.504	2.456	2.452	2.477		
		0.5	2.591	2.593	2.610	2.363	2.265	2.362	2.318	2.402	2.438	2.424	2.423	2.438		
	easy	0	2.535	2.562	2.585	2.300	2.242	2.314	2.258	2.480	2.537	2.489	2.533	2.576		
		0.5	2.582	2.590	2.610	2.269	2.220	2.264	2.235	2.420	2.484	2.430	2.505	2.563		

Note. The smallest values among the equating methods are boldfaced. RMSD = root mean square difference.

Table 4. The Overall RMSD for Different Equating Methods (Number of Items = 30).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches										circle-arc	Rasch
				IMP	IMP total	IMP pair	IMP total	IMP pair	IMP total	circle	Tucker	Levine	Equipper-centile		
0.2	50	rand	0	4.055	4.042	4.019	3.110	3.287	3.099	3.118	3.281	3.537	3.339	3.329	3.337
			0.5	3.970	3.952	3.937	3.066	3.201	3.073	3.068	3.243	3.479	3.347	3.271	3.292
		hard	0	4.039	4.025	4.007	3.127	3.257	3.128	3.106	3.317	3.606	3.409	3.320	3.365
			0.5	4.110	4.036	3.971	3.107	3.207	3.110	3.075	3.227	3.500	3.312	3.223	3.289
		easy	0	4.016	3.999	3.970	3.108	3.266	3.095	3.111	3.298	3.399	3.389	3.328	3.391
	0.5		3.977	3.943	3.924	3.060	3.210	3.052	3.061	3.338	3.531	3.407	3.339	3.409	
	0		3.938	3.921	3.931	3.016	3.032	3.007	2.976	3.192	3.301	3.234	3.197	3.233	
	100	rand	0.5	3.961	3.947	3.939	3.075	3.068	3.065	3.068	3.156	3.284	3.199	3.163	3.197
			0	4.049	4.066	4.060	3.152	3.129	3.152	3.107	3.217	3.363	3.259	3.199	3.257
		hard	0.5	4.076	4.059	4.018	3.191	3.113	3.184	3.159	3.177	3.324	3.217	3.171	3.197
0			3.868	3.867	3.887	2.984	3.010	2.992	2.967	3.188	3.306	3.218	3.216	3.266	
easy		0.5	3.911	3.926	3.919	3.002	2.989	2.997	2.964	3.151	3.304	3.183	3.175	3.229	
	0	3.348	3.288	3.249	2.772	2.929	2.776	2.791	3.032	3.147	3.088	3.070	3.051		
	0.5	3.350	3.304	3.277	2.840	2.956	2.846	2.841	3.007	3.085	3.091	3.083	3.038		
0.3	50	rand	0	3.474	3.377	3.380	2.936	3.031	2.940	2.917	3.062	3.227	3.173	3.118	3.061
			0.5	3.440	3.390	3.322	2.941	2.972	2.952	2.874	2.992	3.124	3.071	3.039	3.002
		hard	0	3.330	3.281	3.270	2.770	2.951	2.762	2.827	3.074	3.170	3.170	3.137	3.126
			0.5	3.349	3.314	3.286	2.756	2.937	2.755	2.786	3.040	3.197	3.101	3.097	3.107
		easy	0	3.353	3.335	3.329	2.840	2.888	2.849	2.785	2.991	3.060	3.024	3.003	2.991
	0.5		3.350	3.309	3.299	2.875	2.851	2.865	2.796	2.999	3.074	3.044	3.006	3.008	
	0		3.418	3.391	3.332	2.944	2.887	2.961	2.867	3.014	3.096	3.046	2.998	3.049	
	100	rand	0.5	3.394	3.347	3.297	3.062	2.962	3.070	2.947	2.959	3.024	3.001	2.950	2.991
			0	3.277	3.267	3.243	2.795	2.831	2.780	2.742	3.042	3.075	3.087	3.056	3.079
		hard	0.5	3.373	3.343	3.351	2.789	2.855	2.800	2.769	3.004	3.065	3.038	3.004	3.032
0			3.373	3.343	3.351	2.789	2.855	2.800	2.769	3.004	3.065	3.038	3.004	3.032	
easy		0	3.373	3.343	3.351	2.789	2.855	2.800	2.769	3.004	3.065	3.038	3.004	3.032	

Note. The smallest values among the equating methods are boldfaced. RMSD = root mean square difference.

Table 5. The Overall RMSD for Different Equating Methods (Number of Items = 40).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches										Linear method			
				IMP	IMP total	IMP pair	IMP total_ Tucker	IMP pair_ Tucker	IMP total_ circle	IMP Tucker_ circle	Tucker	Levine	Equipper-centile	circle-arc	Rasch		
0.2	50	rand	0	4.893	4.832	4.741	3.640	3.974	3.649	3.699	3.807	4.063	3.897	3.823	3.920		
			0.5	5.008	4.906	4.807	3.657	3.964	3.642	3.653	3.757	4.013	3.880	3.803	3.885		
		hard	0	4.820	4.729	4.642	3.663	3.909	3.689	3.675	3.686	3.954	3.836	3.792	3.891		
			0.5	4.832	4.760	4.646	3.812	3.904	3.811	3.736	3.648	3.913	3.768	3.775	3.759		
			0	4.839	4.730	4.670	3.592	3.894	3.587	3.638	3.780	4.116	3.902	3.827	3.961		
	100	rand	0	5.030	4.918	4.802	3.655	3.915	3.640	3.685	3.711	3.940	3.824	3.781	3.858		
			0.5	4.876	4.827	4.731	3.571	3.749	3.577	3.558	3.697	3.881	3.749	3.733	3.803		
		hard	0	4.892	4.823	4.703	3.735	3.785	3.726	3.717	3.609	3.760	3.671	3.669	3.727		
			0	4.845	4.803	4.684	3.710	3.751	3.686	3.638	3.621	3.771	3.674	3.704	3.768		
			0.5	4.899	4.803	4.659	3.824	3.824	3.855	3.765	3.577	3.701	3.648	3.627	3.688		
0.3	50	easy	0	4.853	4.827	4.740	3.519	3.690	3.516	3.508	3.665	3.778	3.714	3.708	3.775		
			0.5	4.933	4.889	4.798	3.512	3.644	3.505	3.509	3.603	3.713	3.665	3.648	3.710		
		rand	0	4.010	3.925	3.893	3.232	3.586	3.231	3.401	3.448	3.626	3.574	3.553	3.563		
			0.5	4.073	3.981	3.898	3.327	3.576	3.303	3.403	3.428	3.570	3.556	3.555	3.545		
			0	4.065	3.969	3.943	3.444	3.642	3.438	3.489	3.439	3.562	3.612	3.544	3.584		
	100	easy	0	4.039	3.901	3.797	3.520	3.576	3.510	3.428	3.363	3.476	3.530	3.436	3.460		
			0	4.085	4.023	3.984	3.300	3.656	3.307	3.438	3.524	3.649	3.665	3.631	3.638		
		rand	0	4.118	4.058	3.966	3.249	3.617	3.235	3.394	3.483	3.621	3.633	3.542	3.581		
			0.5	3.935	3.870	3.803	3.317	3.375	3.314	3.280	3.435	3.476	3.491	3.503	3.535		
			0	4.012	3.924	3.888	3.484	3.463	3.517	3.385	3.462	3.561	3.535	3.549	3.611		
100	easy	0	3.916	3.898	3.752	3.547	3.441	3.561	3.414	3.338	3.424	3.407	3.409	3.485			
		0	3.999	3.954	3.891	3.287	3.428	3.309	3.286	3.473	3.537	3.497	3.572	3.599			
	rand	0	4.113	4.063	4.009	3.271	3.419	3.290	3.276	3.462	3.492	3.538	3.539	3.597			

Note. The smallest values among the equating methods are boldfaced. RMSD = root mean square difference.

Table 6. The Overall RMSD for Different Equating Methods (Number of Items = 50).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches										Linear method			
				IMP	IMP_total	IMP_pair	IMP_total	IMP_circle	Tucker	circle	IMP_Tucker_circle	Tucker	Levine	Equipper-centile	circle-arc	Rasch	
																	IMP
0.2	50	rand	0	5.767	5.645	5.477	4.321	4.785	4.323	4.461	4.271	4.489	4.485	4.320	4.275		
			0.5	5.829	5.651	5.438	4.379	4.694	4.353	4.390	4.194	4.410	4.353	4.248	4.220		
		hard	0	5.848	5.709	5.510	4.342	4.807	4.553	4.488	4.488	4.379	4.389	4.339	4.292	4.245	
			0.5	5.807	5.659	5.446	4.537	4.775	4.543	4.521	4.192	4.403	4.433	4.339	4.307	4.245	
			0	5.599	5.499	5.352	4.160	4.627	4.169	4.308	4.211	4.449	4.428	4.428	4.339	4.392	
	100	rand	0	5.801	5.657	5.475	4.251	4.764	4.259	4.403	4.221	4.460	4.416	4.423	4.456		
			0	5.754	5.679	5.497	4.236	4.497	4.242	4.269	4.203	4.333	4.286	4.225	4.240		
		hard	0	5.693	5.605	5.398	4.350	4.511	4.346	4.322	4.185	4.314	4.285	4.173	4.196		
			0	5.830	5.737	5.537	4.426	4.603	4.416	4.415	4.226	4.359	4.334	4.224	4.253		
			0.5	5.761	5.582	5.374	4.622	4.581	4.621	4.516	4.144	4.239	4.232	4.154	4.164		
0.3	50	rand	0	5.688	5.596	5.465	4.209	4.488	4.220	4.270	4.221	4.329	4.326	4.254	4.316		
			0.5	5.775	5.641	5.462	4.202	4.435	4.217	4.247	4.164	4.287	4.239	4.200	4.263		
		hard	0	4.530	4.389	4.408	3.787	4.135	3.777	3.962	3.965	4.092	4.169	4.086	4.075		
			0.5	4.666	4.511	4.420	3.813	4.142	3.832	3.953	3.917	4.041	4.144	4.074	4.055		
			0	4.832	4.700	4.656	4.021	4.341	4.038	4.159	3.945	4.081	4.174	4.085	4.072		
	100	rand	0	4.717	4.572	4.477	4.248	4.237	4.269	4.118	3.841	4.013	4.073	4.048	3.930		
			0.5	4.619	4.534	4.473	3.903	4.221	3.884	4.041	3.994	4.147	4.241	4.161	4.171		
		hard	0	4.710	4.585	4.542	3.817	4.227	3.807	4.008	3.919	4.088	4.174	4.127	4.125		
			0	4.454	4.433	4.424	3.818	4.015	3.829	3.891	3.888	3.943	3.970	3.936	3.976		
			0.5	4.356	4.305	4.306	3.823	3.928	3.821	3.842	3.805	3.830	3.878	3.892	3.883		
easy	rand	0	4.455	4.463	4.412	3.941	4.038	3.956	3.936	3.786	3.850	3.847	3.863	3.869			
		0.5	4.534	4.516	4.356	4.202	4.122	4.155	4.044	3.786	3.819	3.879	3.816	3.870			
	hard	0	4.474	4.436	4.400	3.817	4.023	3.833	3.872	3.887	3.943	3.959	4.029	4.091			
		0.5	4.462	4.383	4.383	3.764	3.945	3.765	3.820	3.847	3.910	3.966	3.986	4.065			

Note. The smallest values among the equating methods are boldfaced.

Table 7. The Averaged BIAS for Different Equating Methods (Number of Items = 20).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches										Linear method			
				IMP		IMP_total		IMP_pair		IMP_total		IMP_Tucker		circle	circle-arc	circle-arc	Rasch
				total	pair	Tucker	total	Tucker	total	Tucker	Tucker	Levine					
0.2	50	rand	0	2.371	2.386	2.402	1.941	1.953	1.935	1.933	2.115	2.328	2.146	2.121	2.160		
			0.5	2.388	2.409	2.397	1.941	1.941	1.937	1.926	2.096	2.338	2.155	2.108	2.142		
		hard	0	2.441	2.445	2.446	1.968	1.977	1.961	1.949	2.121	2.239	2.165	2.130	2.160		
			0.5	2.465	2.472	2.464	1.934	1.950	1.924	1.919	2.093	2.268	2.123	2.082	2.121		
		easy	0	2.359	2.376	2.373	1.944	1.952	1.956	1.942	2.151	2.309	2.168	2.173	2.209		
			0.5	2.393	2.380	2.379	1.923	1.929	1.918	1.921	2.111	2.296	2.143	2.139	2.182		
	100	rand	0	2.419	2.426	2.416	1.932	1.916	1.941	1.928	2.093	2.187	2.099	2.107	2.136		
			0.5	2.417	2.421	2.413	1.926	1.905	1.937	1.923	2.084	2.163	2.097	2.091	2.114		
		hard	0	2.451	2.453	2.450	1.957	1.931	1.952	1.938	2.108	2.231	2.125	2.082	2.122		
			0.5	2.417	2.413	2.417	1.955	1.922	1.962	1.931	2.046	2.125	2.064	2.045	2.074		
		easy	0	2.326	2.335	2.330	1.923	1.896	1.921	1.922	2.098	2.164	2.106	2.124	2.184		
			0.5	2.400	2.404	2.396	1.943	1.923	1.946	1.935	2.108	2.209	2.114	2.130	2.180		
0.3	50	rand	0	2.071	2.069	2.073	1.769	1.791	1.777	1.757	1.962	2.066	1.991	1.993	2.011		
			0.5	2.077	2.081	2.067	1.751	1.779	1.751	1.733	1.958	2.054	1.974	1.976	1.995		
		hard	0	2.079	2.054	2.069	1.807	1.795	1.808	1.782	1.941	2.037	1.955	1.961	1.968		
			0.5	2.058	2.054	2.069	1.814	1.784	1.806	1.765	1.910	2.031	1.928	1.919	1.913		
		easy	0	1.998	2.003	2.019	1.761	1.764	1.754	1.731	1.943	1.978	1.951	2.000	2.041		
			0.5	2.027	2.033	2.019	1.765	1.766	1.767	1.742	1.960	2.000	1.973	2.015	2.057		
	100	rand	0	2.037	2.052	2.060	1.791	1.748	1.781	1.744	1.921	1.974	1.933	1.950	1.965		
			0.5	2.042	2.054	2.064	1.828	1.757	1.830	1.784	1.905	1.935	1.915	1.941	1.948		
		hard	0	2.115	2.116	2.144	1.878	1.814	1.876	1.830	1.940	1.999	1.958	1.956	1.978		
			0.5	2.071	2.076	2.093	1.900	1.822	1.899	1.863	1.919	1.947	1.934	1.930	1.946		
		easy	0	2.039	2.056	2.070	1.841	1.805	1.849	1.812	1.973	2.023	1.982	2.017	2.051		
			0.5	2.081	2.084	2.102	1.815	1.779	1.813	1.790	1.933	1.978	1.943	2.003	2.047		

Note. The smallest values among the equating methods are boldfaced.

Table 8. The Averaged BIAS for Different Equating Methods (Number of Items = 30).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches										Linear method						
				IMP		IMP_pair		IMP_total		IMP_Tucker		IMP_circle		IMP_Tucker_circle		Tucker	Levine	Equipper-centile	circle-arc	Rasch
				total	pair	total	pair	Tucker	circle	Tucker	circle	Tucker	circle	Tucker	Levine	Equipper-centile	circle-arc	Rasch		
0.2	50	rand	0	3.278	3.257	3.246	2.507	2.653	2.496	2.511	2.617	2.807	2.663	2.656	2.664					
			0.5	3.199	3.183	3.162	2.471	2.572	2.475	2.468	2.587	2.759	2.656	2.614	2.622					
		hard	0	3.264	3.257	3.243	2.519	2.632	2.517	2.502	2.644	2.852	2.709	2.643	2.682					
			0.5	3.294	3.242	3.197	2.514	2.582	2.513	2.479	2.570	2.766	2.629	2.577	2.625					
			easy	0	3.247	3.240	3.210	2.495	2.635	2.489	2.613	2.683	2.689	2.638	2.688					
	100	rand	0	3.188	3.166	3.143	2.463	2.575	2.464	2.460	2.677	2.810	2.727	2.680	2.737					
			0.5	3.168	3.157	3.164	2.427	2.440	2.421	2.394	2.539	2.620	2.572	2.549	2.576					
		hard	0	3.176	3.163	3.161	2.477	2.465	2.467	2.467	2.517	2.614	2.548	2.524	2.545					
			0	3.230	3.258	3.263	2.546	2.518	2.544	2.506	2.564	2.676	2.599	2.552	2.598					
			0.5	3.253	3.239	3.208	2.580	2.519	2.568	2.560	2.538	2.654	2.571	2.529	2.554					
0.3	50	rand	0	3.102	3.098	3.108	2.402	2.420	2.411	2.397	2.544	2.633	2.568	2.608						
			0.5	3.132	3.145	3.137	2.427	2.408	2.416	2.394	2.512	2.626	2.541	2.532	2.573					
		hard	0	2.689	2.638	2.607	2.223	2.351	2.234	2.240	2.412	2.508	2.459	2.445	2.428					
			0.5	2.706	2.673	2.653	2.301	2.392	2.302	2.297	2.415	2.474	2.475	2.474	2.440					
			easy	0	2.794	2.723	2.732	2.372	2.445	2.375	2.354	2.445	2.568	2.524	2.483	2.439				
	100	rand	0	2.759	2.715	2.682	2.375	2.398	2.377	2.309	2.383	2.486	2.440	2.423	2.390					
			0.5	2.677	2.627	2.624	2.224	2.367	2.217	2.267	2.447	2.528	2.522	2.496	2.494					
		hard	0	2.694	2.671	2.645	2.219	2.365	2.225	2.253	2.430	2.554	2.480	2.474	2.482					
			0	2.702	2.680	2.681	2.290	2.325	2.285	2.241	2.386	2.442	2.416	2.400	2.386					
			0.5	2.692	2.654	2.658	2.305	2.295	2.301	2.253	2.384	2.441	2.409	2.389	2.384					
easy	rand	0	2.737	2.712	2.684	2.372	2.336	2.385	2.323	2.411	2.474	2.431	2.391	2.436						
		0.5	2.709	2.671	2.650	2.491	2.405	2.503	2.404	2.353	2.405	2.378	2.342	2.372						
	hard	0	2.639	2.625	2.610	2.246	2.280	2.231	2.213	2.423	2.452	2.457	2.433	2.452						
		0.5	2.720	2.690	2.697	2.247	2.310	2.249	2.241	2.405	2.448	2.434	2.402	2.426						

Note. The smallest values among the equating methods are boldfaced.

Table 9. The Averaged BIAS for Different Equating Methods (Number of Items = 40).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches										Linear method				
				IMP		IMP_total_		IMP_pair_		IMP_total_		IMP_Tucker_		Tucker	Levine	Equipper-centile	circle-arc	Rasch
				total	pair	Tucker	total	Tucker	total	circle	circle	circle						
0.2	50	rand	0	3.946	3.893	3.818	2.910	3.191	2.927	2.957	3.031	3.214	3.111	3.039	3.118			
			0.5	3.980	3.914	3.832	2.951	3.167	2.943	2.928	2.995	3.189	3.101	3.033	3.108			
		hard	0	3.902	3.814	3.757	2.965	3.169	2.989	2.975	2.965	3.163	3.064	3.050	3.121			
			0.5	3.881	3.822	3.748	3.098	3.158	3.090	3.024	2.929	3.096	3.020	3.023	3.011			
			0	3.916	3.819	3.777	2.891	3.148	2.885	2.937	3.023	3.265	3.125	3.051	3.164			
	100	rand	0	4.042	3.946	3.854	2.937	3.142	2.927	2.964	2.967	3.136	3.056	3.022	3.083			
			0.5	3.897	3.862	3.789	2.882	3.024	2.878	2.873	2.953	3.102	3.002	2.978	3.036			
		hard	0	3.903	3.856	3.760	3.026	3.058	3.017	2.944	2.882	3.000	2.930	2.947	3.010			
			0.5	3.887	3.846	3.759	3.001	3.038	2.978	2.944	2.856	2.956	2.906	2.892	2.944			
			0	3.896	3.832	3.734	3.129	3.112	3.142	3.073	2.828	2.922	3.011	2.927	3.012			
0.3	50	rand	0	3.897	3.871	3.803	2.837	2.959	2.818	2.831	2.880	2.961	2.961	2.920	2.968			
			0.5	3.954	3.918	3.843	2.829	2.933	2.818	2.886	2.599	2.742	2.759	2.879	2.852			
		hard	0	3.232	3.157	3.131	2.601	2.886	2.599	2.881	2.681	2.746	2.748	2.848	2.834			
			0.5	3.272	3.193	3.139	2.696	2.881	2.681	2.746	2.681	2.746	2.748	2.848	2.841			
			0	3.251	3.168	3.160	2.772	2.927	2.765	2.882	2.841	2.807	2.750	2.843	2.884			
	100	rand	0	3.234	3.114	3.058	2.853	2.882	2.882	2.841	2.761	2.775	2.791	2.727	2.745			
			0.5	3.273	3.220	3.194	2.658	2.931	2.660	2.763	2.815	2.920	2.933	2.884	2.902			
		hard	0	3.308	3.254	3.193	2.610	2.909	2.604	2.733	2.781	2.888	2.889	2.830	2.864			
			0	3.117	3.039	3.072	2.617	2.699	2.611	2.589	2.749	2.783	2.790	2.803	2.830			
			0.5	3.156	3.096	3.042	2.680	2.715	2.674	2.646	2.687	2.746	2.751	2.750	2.785			
easy	rand	0	3.217	3.150	3.134	2.802	2.800	2.833	2.738	2.752	2.824	2.808	2.820	2.872				
		0.5	3.148	3.124	3.017	2.896	2.801	2.915	2.784	2.665	2.730	2.717	2.718	2.782				
	hard	0	3.210	3.163	3.124	2.638	2.754	2.656	2.642	2.770	2.825	2.782	2.851	2.875				
		0.5	3.311	3.264	3.229	2.629	2.766	2.651	2.647	2.769	2.790	2.831	2.836	2.876				
		0	3.311	3.264	3.229	2.629	2.766	2.651	2.647	2.769	2.790	2.831	2.836	2.876				

Note. The smallest values among the equating methods are boldfaced.

Table 10. The Averaged BIAS for Different Equating Methods (Number of Items = 50).

Anchor ratio	Sample size	Anchor type	Group difference	Imputation approaches											
				IMPa				IMPb				Linear method			
				IMP total	IMP pair	IMP total	IMP pair	Tucker	circle	IMP total	circle	IMP Tucker	Tucker	Levine	Equipper-centile
0.2	50	rand	0	4.643	4.527	4.421	3.476	3.865	3.492	3.609	3.418	3.582	3.590	3.453	3.427
			0.5	4.663	4.520	4.373	3.556	3.783	3.536	3.552	3.334	3.502	3.483	3.378	3.362
		hard	0	4.677	4.572	4.439	3.481	3.860	3.492	3.610	3.319	3.459	3.483	3.412	3.365
			0.5	4.592	4.477	4.354	3.699	3.828	3.705	3.637	3.317	3.478	3.496	3.403	3.375
			easy	0	4.457	4.387	4.284	3.335	3.704	3.334	3.449	3.342	3.527	3.518	3.462
	100	rand	0	4.655	4.530	4.407	3.430	3.837	3.435	3.551	3.363	3.545	3.531	3.526	3.551
			0.5	4.607	4.547	4.416	3.420	3.635	3.430	3.449	3.340	3.438	3.408	3.360	3.374
		hard	0	4.556	4.478	4.321	3.525	3.649	3.519	3.501	3.332	3.431	3.405	3.318	3.342
			0.5	4.646	4.581	4.439	3.596	3.735	3.577	3.594	3.375	3.483	3.459	3.374	3.389
			easy	0	4.588	4.457	4.302	3.771	3.734	3.768	3.690	3.287	3.369	3.362	3.292
0.3	50	rand	0	4.548	4.468	4.364	3.384	3.610	3.394	3.441	3.362	3.446	3.449	3.401	3.439
			0.5	4.622	4.504	4.385	3.386	3.574	3.398	3.420	3.325	3.419	3.386	3.357	3.405
		hard	0	3.648	3.534	3.564	3.062	3.345	3.049	3.201	3.151	3.258	3.309	3.258	3.245
			0.5	3.735	3.596	3.553	3.078	3.337	3.098	3.193	3.108	3.200	3.280	3.232	3.225
			easy	0	3.882	3.774	3.774	3.271	3.523	3.289	3.381	3.147	3.244	3.332	3.267
	100	rand	0	3.763	3.658	3.619	3.486	3.429	3.503	3.348	3.054	3.177	3.256	3.221	3.118
			0.5	3.728	3.644	3.601	3.140	3.402	3.126	3.260	3.199	3.314	3.392	3.350	3.355
		hard	0	3.789	3.685	3.671	3.078	3.416	3.067	3.236	3.124	3.251	3.314	3.301	3.285
			0.5	3.588	3.568	3.570	3.083	3.248	3.085	3.148	3.115	3.160	3.179	3.150	3.188
			easy	0	3.517	3.471	3.464	3.099	3.183	3.098	3.112	3.034	3.053	3.092	3.104
100	rand	0	3.595	3.610	3.553	3.203	3.282	3.214	3.206	3.015	3.061	3.072	3.080	3.081	
		0.5	3.685	3.652	3.524	3.424	3.366	3.391	3.307	2.987	3.015	3.063	3.002	3.043	
	hard	0	3.588	3.561	3.519	3.070	3.243	3.082	3.132	3.086	3.137	3.148	3.189	3.255	
		0.5	3.601	3.533	3.531	3.034	3.191	3.035	3.087	3.071	3.130	3.168	3.188	3.256	

Note. The smallest values among the equating methods are boldfaced.

Ratio of Anchor Items

As the ratio of anchor items increased, the equating accuracy of all methods improved as their RMSD and BIAS values dropped without exception. In particular, the RMSD and BIAS values of the three methods using only raw data or their integrated information (IMP, IMP_total, and IMP_pair method) declined the most as the ratio of anchor items increased.

Ability Distribution

The equating accuracy of all equating methods was very similar regardless of the ability distribution of the second group. That said, the difference in the mean ability of the two groups does not affect the performance of the reference methods in terms of equating accuracy.

Anchor Type

The gaps in RMSD and BIAS values among the three anchor types for all equating methods were not substantial. Their RMSD and BIAS values for imputation methods were slightly larger when hard anchor items were used than the random or the easy sets.

Conclusion and Discussion

Based on different data augmentation methods, seven CRF-based imputation methods were proposed to perform equating in a NEAT design. The performance of seven imputation methods and several traditional equating methods was investigated under varying sample sizes, test lengths, the ratios of anchor items, the difference in examinee ability, and the types of anchors. The findings suggest that imputation methods incorporated with the wisdom of other methods (e.g., the Tucker method or circle-arc method) yield the highest equating accuracy when the test length is short, but when the test length reaches 50, the Tucker method shows a slight advantage. Increasing the sample size does not always reduce equating errors for the proposed methods; this finding makes the largest difference between the imputation methods and the reference ones. Furthermore, the lower the proportion of anchor items, the worse the performance of all equating methods, while the type of anchor items and group ability differences had little impact on the equating results.

The imputation methods possess high flexibility in subsuming good results from various augmenting strategies and equating methods. Some specific methods from the former (i.e., using different data augmentation methods) can be unstable, especially for the one using response data of the anchor test only, leading to low equating accuracy. On the other hand, the latter kind that combines information from other equating methods' results can significantly improve the performance, even better than the original ones that were selected for augmentations. Particularly, given the Tucker

method is selected to incorporate into the proposed methods, the equating accuracy obtained by using the total scores of the anchor test (IMP_total_Tucker) is better than using the sum scores of anchor item pairs (IMP_pair_Tucker) in most cases. IMP_pair_Tucker method performs better only when the test length is the shortest (20) and the sample size is relatively large (100). In addition, IMP_Tucker_cirlce method which uses the most information is more advantageous in short tests.

Therefore, we suggest that when the test length is not more than 40, imputation methods with more information to augment the data set, such as aggregated scores from the test itself and information from other equating methods (e.g., IMP_Tucker_cirlce), are recommended. Moreover, when the test length is extremely short, say less than 20, and the sample size is relatively large, IMP_pair_Tucker method is also applicable.

In this study, we not only set a small sample situation but also limited the test length to a short range (50 items and below) to be suitable for equating the analysis of short tests. Short tests are very popular in educational measurement, such as quizzes, unit tests, and subtests in comprehensive tests. There have been some equating studies on short tests that contain 40 or fewer items (Dimitrov, 2018; Lim & Lee, 2020), but few studies have focused on small samples equating with short tests, although such scenarios are not rare in educational practice. While comparing the performance of various equating methods under this condition, this study proposes several imputation methods that are particularly suitable for this case. The imputation methods can also be directly extended to polytomous scoring situations for equating mental health questionnaires using Likert-type scales, which are usually short in length.

Although the current research successfully used the ML-based imputation technique to perform equating tasks within the NEAT design in the small sample scenario, several limitations should be considered in future studies. (a) As the proposed methods were developed and applied for dichotomous items only, future research can extend the application to polytomous or mixed-format cases. (b) The proposed methods were based on the CRF-based method; other augmentation strategies, such as adding group Y's total scores on test form#2 into the data, can be considered in the equating research studies. (c) Unidimensionality and local independency are the main assumptions of IRT models. In some psychological or educational tests, unidimensionality may not be fully satisfied, or local dependency usually exists in practice. Multidimensional equating methods or testlet-based equating methods can be considered to treat multidimensional measures or address local dependence between items in future research. (d) Although the evaluation criteria used in most equating research studies were based on the recovery of the true values, the evaluation criteria of equating errors have always been a difficulty in equating research studies. Determining or finding consistent evaluation criteria or "gold standard" in equating research studies deserves further investigation.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Z.J. was supported by the National Natural Science Foundation of China for Young Scholars (grant no. 72104006) and Peking University Health Science Center (grant no. BMU2021YJ010).

ORCID iDs

Lingling Xu  <https://orcid.org/0000-0001-7112-2134>

Dexin Shi  <https://orcid.org/0000-0002-4120-6756>

Ren Liu  <https://orcid.org/0000-0002-6708-4996>

Supplemental Material

Supplementary material for this article is available online.

References

- Albano, A. D. (2016). Equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, 74, 1–36.
- Arai, S., & Mayekawa, S. I. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, 38(1), 1–16.
- Athey, S. (2018). The impact of machine learning on economics. *The Economics of Artificial Intelligence: An Agenda* (pp. 507–547). University of Chicago Press.
- Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement*, 72, 608–628. <https://doi.org/10.1177/0013164411428609>
- Babcock, B., & Hodge, K. (2020). Rasch versus classical equating in the context of small sample sizes. *Educational and Psychological Measurement*, 80(3), 499–521. <https://doi.org/10.1177/0013164419878483>
- Battaui, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, 68, 1–22.
- Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press.
- Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Addison-Wesley.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Dimitrov, D. M. (2018). The delta-scoring method of tests with binary items: A note on true score estimation and equating. *Educational and Psychological Measurement*, 78(5), 805–825. <https://doi.org/10.1177/0013164417724187>
- Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement*, 53(1), 3–22. <https://doi.org/10.1111/jedm.12098>
- Enders, C. K. (2010). *Applied missing data analysis*. The Guilford press.

- Equihua, M. (1990). Fuzzy clustering of ecological data. *Journal of Ecology*, *78*, 519–525.
- González, J. (2014). SNSequate: Standard and nonstandard statistical models and methods for test equating. *Journal of Statistical Software*, *59*, 1–30.
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24. <https://doi.org/10.1177/0146621602026001001>
- Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183. <https://doi.org/10.3102/10769986025002133>
- Hong, S., Sun, Y., Li, H., & Lynn, H. S. (2020). *Influence of parallel computing strategies of iterative imputation of missing data: A case study on missForest*. arxiv Preprint arxiv:2004.11195.
- Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of irt-based equating methods in the presence of outlier common items. *Applied Psychological Measurement*, *32*(4), 311–333. <https://doi.org/10.1177/0146621606292215>
- Ij, H. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234.
- Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review*, *13*(2), 311–321.
- Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131–143. <https://doi.org/10.1177/01466216980222003>
- Kim, S. H., Von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement*, *45*(4), 325–342. <https://doi.org/10.1111/j.1745-3984.2008.00068.x>
- Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practice* (2nd ed.). Springer.
- Lakshminarayan, K., Harp, S. A., Goldman, R. P., & Samad, T. (1996, August). *Imputation of missing data using machine learning techniques*. KDD-96 Proceedings. <https://www.aaai.org/Papers/KDD/1996/KDD96-023.pdf>
- Lim, E., & Lee, W.-C. (2020). Subscore equating and profile reporting. *Applied Measurement in Education*, *33*, 295–112. <https://doi.org/10.1080/08957347.2020.1732381>
- Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*(2), 1487–1509.
- Liou, M., & Cheng, P. E. (1995). Equipercntile equating via data-imputation techniques. *Psychometrika*, *60*(1), 119–136.
- Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, *25*, 197–207. <https://doi.org/10.1177/01466210122032000>
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data (Vol. 793)*. Wiley.
- 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>
- Maris, G., Schmittmann, V. D., & Borsboom, D. (2010). Who needs linear equating under the NEAT design? *Measurement: Interdisciplinary Research & Perspective*, *8*(1), 11–15. <https://doi.org/10.1080/15366361003684653>

- Mayer, M., & Mayer, M. M. (2022, April 28). *Package "missRanger"* [R package]. <https://cran.hafro.is/web/packages/missRanger/missRanger.pdf>
- Moses, T., Deng, W., & Zhang, Y. L. (2011). Two approaches for using multiple anchors in NEAT equating: A description and demonstration. *Applied Psychological Measurement, 35*(5), 362–379. <https://doi.org/10.1177/0146621611405510>
- Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., & . . . Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution, 5*(9), 961–970. <https://doi.org/10.1111/2041-210X.12232>
- Perry, R. P., & Dickens, W. J. (1987). Perceived control and instruction in the college classroom: Some implications for student achievement. *Research in Higher Education, 27*(4), 291–310.
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.Rproject.org/>
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology, 179*(6), 764–774.
- Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249–275. <https://doi.org/10.1111/j.1745-3984.2007.00037.x>
- Sinharay, S., & Holland, P. W. (2010). The missing data assumptions of the NEAT design and their implications for test equating. *Psychometrika, 75*(2), 309–327.
- Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309–330. <https://doi.org/10.1111/j.1745-3984.2005.00018.x>
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
- Stewart, J., & Gibson, A. (2010). Equating classroom pre and post tests under item response theory. *Shiken: JALT Testing and Evaluation SIG Newsletter, 14*(2), 11–18.
- Von Davier, A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of equating*. Springer.
- Wang, T., Lee, W. C., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632–651. <https://doi.org/10.1177/0146621608314943>
- Wolkowitz, A. A., & Wright, K. D. (2019). Effectiveness of equating at the passing score for exams with small sample sizes. *Journal of Educational Measurement, 56*(2), 361–390.
- Wong, K. C. Y., Xiang, Y., Yin, L., & So, H. C. (2021). Uncovering clinical risk factors and predicting severe COVID-19 cases using UK biobank data: Machine learning approach. *JMIR Public Health and Surveillance, 7*(9), Article e29544. <https://doi.org/10.2196/29544>
- Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems, 160*, 104–118. <https://doi.org/10.1016/j.knosys.2018.06.012>
- Zeng, L. (1993). A numerical approach for computing standard errors of linear equating. *Applied Psychological Measurement, 17*(2), 177–186.