METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

# Empirical Ensemble Equating Under the NEAT Design Inspired by Machine Learning Ideology

Zhehan Jiang [1,2], Yuting Han [1,2§], Jihong Zhang [3§], Lingling Xu [1,2], Dexin Shi [4], Haiying Liang [5], Jinying Ouyang [1,2]

**[1]** *Institute of Medical Education, Peking University, Beijing, China.* **[2]** *National Center for Health Professions Education Development, Peking University, Beijing, China.* **[3]** *College of Education and Health Professions, University of Arkansas, Fayetteville, AR, USA.* **[4]** *Department of Psychology, University of South Carolina, Columbia, SC, USA.* **[5]** *Institute of Education, University College London, London, United Kingdom.*

[§]*These authors contributed equally to this work.*

Methodology, 2023, Vol. 19(2), 116–132, https://doi.org/10.5964/meth.10371

**Received:** 2022-09-28 • **Accepted:** 2023-03-31 • **Published (VoR):** 2023-06-30

**Handling Editor:** Katrijn van Deun, Tilburg University, Tilburg, The Netherlands

**Corresponding Author:** Yuting Han, Peking University Health Science Center, 38 Xueyuan Rd, Haidian District, 100191, Beijing, China. E-mail: hanyuting@bjmu.edu.cn

## Abstract

This study proposes an empirical ensemble equating (3E) approach that collectively selects, adopts, weighs, and combines outputs from different sources to take and combine advantage of equating techniques in various score intervals. The ensemble idea was demonstrated and tailored to the Non-Equivalent groups with Anchor Test (NEAT) equating. A simulation study based on several published settings was conducted. Three outcome measures – average bias, its absolute value, and root mean square difference – were used to evaluate the selected methods' performance. The 3E approach outperformed other counterparts in most given conditions, while the cautions, such as tuning weights and assuming possible scenarios for using the proposed approach were also addressed.

## Keywords

ensemble learning, equating, machine learning, NEAT, educational assessment

In high-stakes assessments (e.g., licensure and certification exams), new forms are typically created for continuing test administration. Using a new form at each administration enhances content security and item-exposure control and supports computerized

mechanisms and item bank construction. From a measurement perspective, different assessment forms should be built on an identical set of content and statistical specifications for consistency purposes. Further, statistical models are adopted to support the exchangeability of scores across the forms; this process is generally called equating, allowing computations of scores projected from one form to the other.

Among many equating designs, the Non-Equivalent groups with Anchor Test (NEAT) is a highly, if not the most, popular one widely adopted in research and practice. In an application of the NEAT design, a new test $\boldsymbol{x}$ form is equated to an old test $\boldsymbol{y}$ form, a sample takes $\boldsymbol{x}$ from Group X, and a sample takes $\boldsymbol{y}$ from Group Y. In addition, an anchor test is taken by both groups and allows one to study the difference in ability between Group X and Group Y. Group X's true response data on $\boldsymbol{y}$ form and Group Y's true response data on $\boldsymbol{x}$ form are not observable, as they do not actually happen in the administration; this makes the quality evaluation of equating difficult, as no true values are available for the comparative purpose. Therefore, most studies investigating the performance of equating methods are simulation-based (e.g., Andersson & Wiberg, 2017; Moses & Holland, 2010; Sinharay & Holland, 2010). That is, researchers provide empirical conditions to find if specific methods yield better results than other counterparts; the findings are then used to assist method selections.

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, local equating (van der Linden, 2011), Levine nonlinear method, Kernel equating (KE), and others (see Kolen & Brennan, 2004 for details). Specifically, a post-stratification (PSE), Levine observed-score linear, and chained equating (CE) methods are typically used in KE when the NEAT design is present (von Davier et al., 2004). However, these techniques are not consistently performing better than others. In fact, the performance depends on the settings of actual tasks and different score ranges. For instance, Livingston and Kim (2009) found that differences between equating methods in accuracy were small for raw scores near the median of the distribution but large for scores far from the median, and the circle-arc method had higher accuracy in the upper and lower tails of the score distribution compared to mean equating in small samples. Kim and Livingston (2010) show that, in small sample scenarios, CE produced the most reliable results for low scores, while circle-arc ones were better choices in the upper half of the score distribution.

## Ensemble Learning

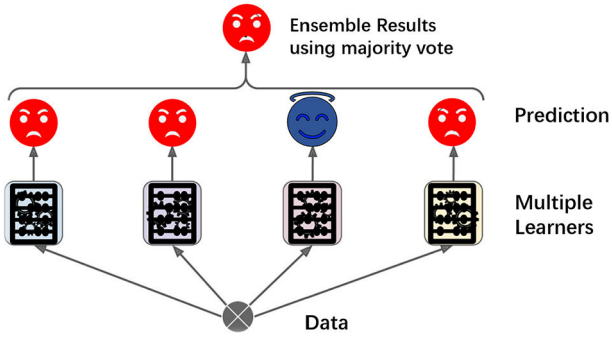
As a powerful technique, ensemble learning (EL) functions like its name suggests: utilizing multiple models to improve the reliability and accuracy of specific machine-learning predictions. The idea of the ensemble is collectively selecting, adopting, weighing, and combining outputs from different sources. Without loss of generality, in a classification

task, techniques such as logistic regression, support vector machine, random forests, and neural network are all set into EL to improve the stability of the overall performance. Tremendous studies across fields show that EL is frequently more reliable than individual models. For instance, Borovkova and Tsiamas (2019) classify different companies in the stock market via EL; Lessmann and colleagues (2021) propose an EL framework to support marketing decision-making; Priore and colleagues (2018) construct scheduling of flexible manufacturing systems using ensemble methods. Unsurprisingly, EL often ranks at the top in machine-learning competitions such as Kaggle (Kumar & Mayank, 2020; Stamp et al., 2021).

EL can be framed in multiple ways. The simplest one is averaging the outputs of different models, while complex ones devising weights and adaptive algorithms to empower the engineer. The concept of EL has been extended to a broader sense, meaning it is not limited to models, but also data and hypotheses. In this paper, we limit EL in the context of a modeling ensemble, where each model is termed a "learner". There are two EL sub-types: sequential EL and parallel EL. The former considers the dependence between learners, each of which is exploited sequentially to obtain more accurate predictions. To illustrate with a classification example again, mislabeled cases have their weights adjusted while the weights for properly labeled sets stay unchanged. Each time a new learner is generated, the weights are updated to improve the classification performance. On the other hand, parallel EL drives learners in parallel. When rendering parallel EL, the idea is to exploit the learners' independence, as the overall error rates can be reduced by drawing on "good" learners' strengths and offsetting "bad" learners' weaknesses. Figure 1 shows a simple EL: four learners (i.e., classification techniques such as logistic regression, support vector machine, and so on) are used to predict a binary variable with a value of red or blue, while the third one yields a different label (blue) to others (red). If one uses majority vote as the ensemble schema, the aggregated result is colored red, as three learners endorse red and only one endorses blue.

**Figure 1**

*A Simple Ensemble Learning Using Majority Vote*



It's self-evident that different weighting schemes can lead to unidentical conclusions, even if the learners are identical in two EL models. "Simple weighted average (SWA)" is that the weights are proportional to the precisions of each learner. "Weight proportional to the square of the precision (SqrWA)" squares the precisions to obtain weights. In contrast, "weight proportional to the precision's powers of $N$ (PrWA)" further extends square to an arbitrary integer $N$. Other schemes, such as considering data collection time (i.e., "age" of the data) and polynomial functions on data variance, are also available but not applicable to the present study (see Wagner, 1975, p. 289). Let's consider a situation where the precisions are wrapped into values larger than 1 (the inversed effect exhibits when the precisions are presented as ratios or percentages); the three weighting schemes (SWA, SqrWA, and PrWA) incrementally entrust the learners that perform the best at a specific estimate more; for example, the same precision will be given more considerable weight in SqrWA than in SWA. An extreme choice is brutally picking the best one and neglecting others; that said, all non-optimal learners receive zeros and 100% for the optimal one when calculating weights.

EL has been applied to different areas and inquiries in educational and psychological studies. Ragab and colleagues (2021) use EL algorithms to predict student failure and enable customized educational paths; Abidi and colleagues (2020) adopt ensemble classifiers to quantify academic procrastination through big data assimilation; Premalatha and Sujatha (2021) predict the employment status of graduates in higher educational institutions via EL; Pearson and colleagues (2019) estimate treatment outcomes following an internet intervention for depression through a machine learning ensemble. These successful applications primarily lie in prediction and classification; engrafting EL to equating tasks remains unknown such that the topic *per se* is practically beneficial and methodologically meaningful to the field.

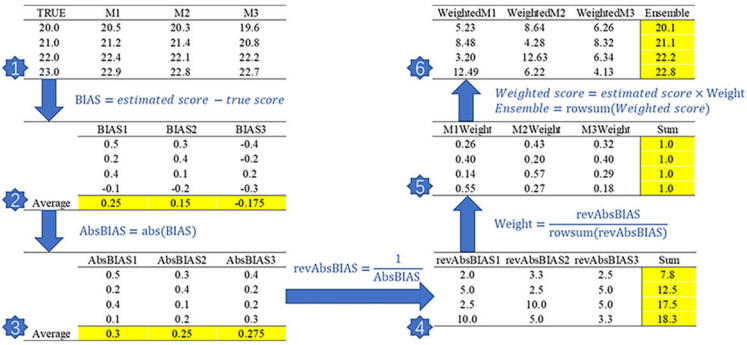
## Method

The method section outlines the steps involved in constructing the proposed ensemble approach and highlights the rationales and consequences of this method through a walkthrough case study. This case study uses the scenario depicted in Figure 2 for illustration purposes. The first part of Figure 2 displays the true scores ranging from 20 to 23, along with the estimated scores generated by three equating models (referred to as learners in this study) represented by M1, M2, and M3. The second and third parts of Figure 2 present the biases (i.e., the equating result minus the true score) and their absolute values, along with their averaged values highlighted in the last row. It can be observed that the lowest values of the averaged bias and absolute bias are 0.15 and 0.25,

respectively (see the last row in the second and third part of Figure 2). Thus, a better approach would ideally produce values lower than these two numbers.

**Figure 2**

*The Steps of a Walkthrough Case Using Simple Weighted Average Schema*



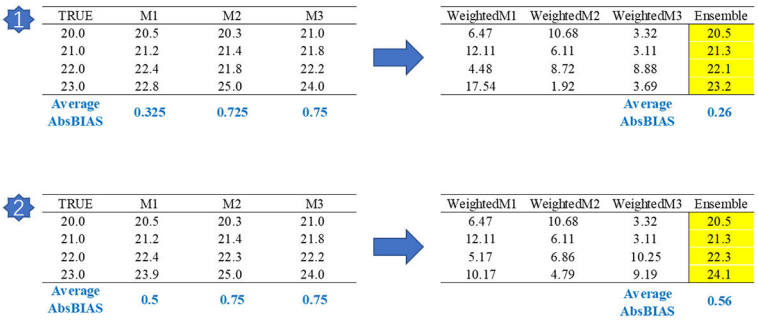

The proposed approach employs a "simple weighted average" schema and requires absolute bias values to generate comparable measures, enabling the calculation of relative contributions to ensemble weights. Theoretically, models with smaller absolute biases should be trusted more and given greater weight in the final ensemble. Thus, contributions to the weights should be inversely related to absolute biases. In the fourth part of Figure 2, the inversions of the absolute biases (e.g., 1.0/0.5, 1.0/0.3, and 1.0/0.4 in the first row) are calculated and summed across each row. These inversions, as shown in the fourth part of Figure 2, are divided by their sums for each row (e.g., 2.0/7.8, 3.3/7.8, and 2.5/7.8 in the first row) to create ensemble weights, which are listed in the fifth part of Figure 2. Consequently, the sum of the weights in each row equals 1, as shown in the last column.

Finally, the weights are applied to the corresponding estimated scores presented in the first part of Figure 2. The ensemble equating is completed by summing the weighted scores, resulting in the last column in the sixth part of Figure 2. It is straightforward to calculate that the average bias and absolute bias values of the ensemble score, as seen in Figure 2 (across the case's 20-23 range), are 0.05 and 0.15. As expected, these aggregated accuracy measures are both lower than those of any individual model, as shown in the last row of the second and third parts of Figure 2. This indicates that the ensemble approach effectively improves the equating accuracy compared to relying on a single model, thus validating the proposed method.

The walkthrough case in Figure 2 shows one scenario only, of which the result is not comprehensive enough to generalize. Figure 3 contains two more scenarios: the first one further amplifies the advantage, as all the aggregated absolute bias values for the

**Figure 3**

*Two Possible Scenarios in the Use of Ensemble Learning*




three learners (0.325, 0.725, and 0.75) are larger than that of the ensemble score (0.26); while the second one, although not producing the optimal results (0.56) when compared with other individual learners (0.5, 0.75, and 0.75), remains robust as it outperforms many counterparts.

As introduced above, different weighting schemes likely result in inconsistent estimates. If one uses the "weight proportional to the square of the value/precision (SqrWA)" scheme, the white cells in the fourth part of Figure 2 should be squared before summing, and the rest of the calculations, follow the same flow. The PrWA_*N* scheme calculates the *N*th power for the reversed absolute bias values in the fourth part of Figure 2 to increase the impact of top-performing learners in the ensemble procedure.

In practice, however, true scores are unknown in an equating setting. Therefore, constructing an ensemble equating model demands a mechanism to account for the plausible variability in observed responses. That said, this mechanism should deliver weights for each learner. Based on the ideology of the walkthrough case, we propose an empirical ensemble equating (3E) approach to handling the NEAT design's inquiry.

Like power analysis in complex scenarios where mathematical deriving fails to provide viable solutions, the 3E approach is simulation-based and rooted in empirical estimates from item response theory (IRT). Let $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{anchor}\}$ be observed responses from a NEAT design and, correspondingly, $\{\beta, \theta\}$ be estimates of item parameters and latent traits via a IRT model (e.g., three-parameter logistic model) where the observed responses are fed. We adopt the famous "KBneat" dataset to demonstrate the 3E approach. This dataset contains responses for two forms (one for each group) of a 36-item NEAT-based examination, while 12 anchor items were taken by both groups (Kolen & Brennan, 2004).

In this study, eight learners were used for both comparative purposes and 3E construction, including linear equating methods (the Tucker linear equating and the

chained linear equating), equipercentile equating methods (the equipercentile equating using frequency estimation method with log-linear smoothing and the equipercentile equating using a chained method with log-linear smoothing), mean equating methods (the Tucker mean equating and the chained mean equating), and the circle-arc ones (the Tucker circle-arc equating and the chained circle-arc equating). These learners have been substantially studied and proved useful in many environments. Employing a diverse set of learners allows the ensemble approach to benefit from collective wisdom, as it can potentially reduce the impact of any shortcomings associated with a single method. By combining the outputs of these learners, the ensemble method can achieve better overall performance, ultimately improving the accuracy and robustness of the equating process.

Each learner was applied to "KBneat", and the equated results were standardized to serve as the ability difference between the two groups: the mean of the standardized differences was 0.3. The "KBneat" data of both groups were calibrated via two separate three-parameter logistic (3-PL) IRT models, resulting in $\beta_X$ and $\beta_Y$ (see Kolen & Brennan, 2014, p. 203 for item parameter estimates). $\theta_X$ is sampled from *Normal* (0, 1), while the $\theta_Y$ is assumed to drift from $\theta_X$ by 0.3 and therefore the distribution was set to *Normal* (0.3, 1.5). $\theta_X$ was used with $\beta_Y$ in Group Y's 3-PL model to generate Group X's *true* responses in *y* form (called ***x.y***), such that each learner's precision was calculated. 10,000 individuals per group were generated as the pool. To summarize, these simulation-based steps were functioning as a basis for learners' weights calculation.

The 3E approach is simulation-based, yet evaluating its performance demands a simulation study, too. Sample size directly affects random equating error, and different equating methods are suitable for different sample sizes. For example, when sample sizes are relatively small, the circle-arc equating and mean equating might be considered (Livingston & Kim, 2009). However, Diao and Keller (2020) suggested that sample sizes between 100 and 400 are sufficient for all classical equating methods. Sample sizes larger than 1000 (e.g., 1000, 1500, 2000) are not well investigated in previous equating research. To fully consider the impact of various sample sizes on the learners (traditional equating methods), the simulation study set the sample size per group equally to [200, 500, 1000, 1500, 2000]. For example, if sample size was 500, 500 rows of the observations were randomly drawn from {***x***, ***y***, ***anchor, x.y***}. As an empirical approach, it's reasonable to assume that item parameters functioned stable. However, the ability difference could vary from cohort to cohort, especially when the sample size was small. Therefore, in addition to 0.3, two more θ drifts—0.1 and 0.5—were used. Finally, the SWA, SqrWA, and PrWA (powers set to 5, 50, and 100) were deployed as weighting schemes of the 3E approach, which adjusted the impact of top-performing learners in the ensemble procedure. Each condition was replicated 100 times. Three measures were used according to the literature (i.e., Wolkowitz & Wright, 2019; Zeng, 1993)—the average bias and its absolute value (BIAS and AbsBIAS) and root mean square difference (RMSD):

$$\text{BIAS} = \frac{\sum_p^{SS}\left(x.y_p - \widehat{x.y_p}\right)}{SS} \tag{1}$$

$$\text{AbsBIAS} = \frac{\sum_p^{SS}\left|x.y_p - \widehat{x.y_p}\right|}{SS} \tag{2}$$

$$\text{RMSD} = \sqrt{\frac{\sum_p^{SS}\left(x.y_p - \widehat{x.y_p}\right)^2}{SS}} \tag{3}$$

where $SS$ was the sample size, and $\widehat{x.y_p}$ was the equated score of an individual examinee. The true responses of individual $p$ from group $X$ on test $y$ ($x.y_p$) was generated through the (3E) approach as described previously. Based on the repeated samples, the measures were calculated by averaging over 100 repetitions. The analysis was implemented using R (Version 4.2.2 64-bit; R Core Team, 2016) and the R code is given in the Supplementary Materials.

## Analysis

Since this study aims to explore the performance of the ensemble method and compare the ensemble method with traditional methods, rather than comparing among traditional methods (learners), in each repetition, we choose the learner with the highest equating accuracy as a reference. The reference method does not explicitly refer to a specific method, which may differ across conditions and repetition. Still, it is always the best learner, and the ensemble method is always compared with the best learner.

The average equating errors for the reference method and five 3E approaches utilizing various weighting schemes across different conditions are displayed in Tables 1, 2, and 3. It is important to note that the reference method does not correspond to a specific equating method. Instead, it represents the method that produced the minimum equating error (i.e., the smallest absolute BIAS, RMSD, and BIAS values) among the eight learners in each repetition, and the values were calculated via averaging over all repetitions.
It is crucial to emphasize that these bias measures are used to gauge the equating accuracy, with smaller values indicating higher precision in the equating process. As shown in Table 1, with the increase of the power in the weighting scheme for the 3E approaches, the smaller the absolute BIAS value, the higher the equating accuracy. This trend weakens until the power increases to 50, and the absolute BIAS value may no longer decrease. The difference in the absolute BIAS values among powers 1, 2, 5, and 50 is relatively large. The difference in the absolute BIAS values between powers 50 and 100 is fairly close. Their equating accuracy is higher than that of the reference method, for their absolute BIAS values are smaller than that of the reference method. That is, in

**Table 1**

*The Averaged Absolute BIAS for Different Equating Methods*

| | | Sample Size | | | | |
|---|---|---|---|---|---|---|
| **Ability Difference** | **Equating Method** | **200** | **500** | **1000** | **1500** | **2000** |
| 0.1 | Reference | 2.263 | 2.213 | 2.134 | 2.151 | 2.149 |
| | PrWA_100 | 2.247 | 2.168 | 2.094 | 2.111 | 2.102 |
| | PrWA_50 | 2.244 | 2.171 | 2.096 | 2.113 | 2.106 |
| | PrWA_5 | 2.506 | 2.552 | 2.475 | 2.505 | 2.521 |
| | SqrWA | 2.648 | 2.738 | 2.665 | 2.691 | 2.711 |
| | SWA | 2.708 | 2.815 | 2.744 | 2.766 | 2.785 |
| 0.3 | Reference | 2.501 | 2.449 | 2.366 | 2.368 | 2.365 |
| | PrWA_100 | 2.483 | 2.417 | 2.338 | 2.347 | 2.339 |
| | PrWA_50 | 2.480 | 2.419 | 2.341 | 2.350 | 2.343 |
| | PrWA_5 | 2.683 | 2.744 | 2.652 | 2.675 | 2.692 |
| | SqrWA | 2.811 | 2.907 | 2.811 | 2.831 | 2.850 |
| | SWA | 2.869 | 2.977 | 2.879 | 2.895 | 2.914 |
| 0.5 | Reference | 2.765 | 2.739 | 2.684 | 2.675 | 2.658 |
| | PrWA_100 | 2.746 | 2.694 | 2.645 | 2.644 | 2.627 |
| | PrWA_50 | 2.739 | 2.693 | 2.645 | 2.645 | 2.630 |
| | PrWA_5 | 2.825 | 2.915 | 2.839 | 2.856 | 2.864 |
| | SqrWA | 2.920 | 3.035 | 2.951 | 2.966 | 2.976 |
| | SWA | 2.969 | 3.092 | 3.004 | 3.017 | 3.026 |

*Note.* PrWA_100, PrWA_50 and PrWA_5 represented the 3E approaches with weight proportional to the precision's powers of 100, 50 and 5, respectively.

the practical equating work with similar conditions, selecting the power of 50 can obtain better equating performance than the reference method.

The absolute BIAS values for all methods increased as the ability difference between the two groups increased. It is worth noting that even though 0.3 is the preset ability difference between the two groups—that is, the weights used in the 3E approaches were calculated under the same setting—the equating deviation of all approaches is still greater than that of the condition that the ability difference between the two groups is 0.1. However, the difference between the reference method and the worst-performing 3E approaches decreases with increasing ability drift. Taking the sample size of 200 as an example, when the ability difference was 0.1, 0.3, and 0.5, the absolute BIAS values between the SWA method and the reference method were 0.445, 0.368, and 0.203, respectively.

Regardless of the ability difference, the absolute BIAS values of the reference, the PrWA_100, and PrWA_50 methods decreased as the sample size increased from 200 to

**Table 2**

*The Averaged RMSD for Different Equating Methods*

| | | Sample Size | | | | |
|---|---|---|---|---|---|---|
| **Ability Difference** | **Equating Method** | **200** | **500** | **1000** | **1500** | **2000** |
| 0.1 | Reference | 2.810 | 2.787 | 2.722 | 2.728 | 2.719 |
| | PrWA_100 | 2.801 | 2.735 | 2.675 | 2.687 | 2.670 |
| | PrWA_50 | 2.799 | 2.738 | 2.678 | 2.691 | 2.675 |
| | PrWA_5 | 3.111 | 3.189 | 3.107 | 3.132 | 3.137 |
| | SqrWA | 3.289 | 3.403 | 3.315 | 3.333 | 3.339 |
| | SWA | 3.362 | 3.485 | 3.396 | 3.410 | 3.416 |
| 0.3 | Reference | 3.083 | 3.065 | 2.983 | 2.982 | 2.966 |
| | PrWA_100 | 3.078 | 3.031 | 2.959 | 2.963 | 2.943 |
| | PrWA_50 | 3.076 | 3.033 | 2.961 | 2.966 | 2.947 |
| | PrWA_5 | 3.316 | 3.406 | 3.298 | 3.317 | 3.321 |
| | SqrWA | 3.466 | 3.583 | 3.468 | 3.481 | 3.485 |
| | SWA | 3.532 | 3.655 | 3.538 | 3.548 | 3.550 |
| 0.5 | Reference | 3.369 | 3.385 | 3.322 | 3.312 | 3.280 |
| | PrWA_100 | 3.377 | 3.346 | 3.293 | 3.288 | 3.255 |
| | PrWA_50 | 3.371 | 3.346 | 3.293 | 3.289 | 3.257 |
| | PrWA_5 | 3.476 | 3.587 | 3.492 | 3.506 | 3.499 |
| | SqrWA | 3.580 | 3.717 | 3.609 | 3.621 | 3.614 |
| | SWA | 3.632 | 3.775 | 3.664 | 3.673 | 3.665 |

*Note.* PrWA_100, PrWA_50 and PrWA_5 represented the 3E approaches with weight proportional to the precision's powers of 100, 50 and 5, respectively.

1000. However, larger samples did not always lead to better equating when the sample size was greater than 1000. For example, when the ability difference between the two groups was 0.1 and 0.3, the absolute BIAS values of the reference, the PrWA_100 and the PrWA_50 methods decreased when the sample size increased from 200 to 1000 but increased when the sample size increased from 1000 to 1500. And when the ability difference was 0.5, the absolute BIAS values of the reference, the PrWA_100, and the PrWA_50 methods showed a downward trend with increasing sample size. Larger sample sizes do not always lead to better equating results when the sample size exceeds 1,000 may be due to a phenomenon known as the "law of diminishing returns." As the sample size increases, the estimates derived from the equating methods become more stable and closer to their true values. However, after a certain point, the estimates are already stable enough, and further increasing the sample size provides minimal additional information. Besides, the equating methods may have inherent limitations that prevent them from achieving perfect accuracy, regardless of the sample size. In these cases, increasing the

**Table 3**

*The Averaged BIAS for Different Equating Methods*

| | | Sample Size | | | | |
|---|---|---|---|---|---|---|
| **Ability Difference** | **Equating Method** | **200** | **500** | **1000** | **1500** | **2000** |
| 0.1 | Reference | 0.910 | 0.970 | 0.937 | 0.954 | 0.930 |
| | PrWA_100 | 0.839 | 1.031 | 1.021 | 1.070 | 1.088 |
| | PrWA_50 | 0.849 | 1.044 | 1.031 | 1.081 | 1.100 |
| | PrWA_5 | 1.342 | 1.698 | 1.618 | 1.688 | 1.737 |
| | SqrWA | 1.523 | 1.914 | 1.813 | 1.887 | 1.938 |
| | SWA | 1.592 | 1.990 | 1.881 | 1.955 | 2.007 |
| 0.3 | Reference | 1.560 | 1.602 | 1.538 | 1.537 | 1.507 |
| | PrWA_100 | 1.421 | 1.577 | 1.522 | 1.554 | 1.565 |
| | PrWA_50 | 1.424 | 1.584 | 1.526 | 1.560 | 1.573 |
| | PrWA_5 | 1.729 | 2.044 | 1.924 | 1.981 | 2.024 |
| | SqrWA | 1.857 | 2.205 | 2.065 | 2.126 | 2.172 |
| | SWA | 1.911 | 2.265 | 2.117 | 2.180 | 2.226 |
| 0.5 | Reference | 2.053 | 2.152 | 2.122 | 2.107 | 2.058 |
| | PrWA_100 | 1.880 | 2.043 | 2.022 | 2.032 | 2.022 |
| | PrWA_50 | 1.878 | 2.046 | 2.021 | 2.033 | 2.024 |
| | PrWA_5 | 1.986 | 2.311 | 2.227 | 2.266 | 2.287 |
| | SqrWA | 2.057 | 2.415 | 2.310 | 2.356 | 2.382 |
| | SWA | 2.096 | 2.460 | 2.346 | 2.395 | 2.421 |

*Note.* PrWA_100, PrWA_50 and PrWA_5 represented the 3E approaches with weight proportional to the precision's powers of 100, 50 and 5, respectively.

sample size may not lead to significant improvements in equating accuracy, as the limitations are related to the methods rather than the sample size.

In addition, the effect of sample size on the absolute BIAS values of the PrWA_5, SqrWA and SWA methods did not show a uniform pattern. The RMSD results present a similar pattern to the absolute BIAS values and are shown in Table 2.

Table 3 shows that regardless of which equating method was used, the BIAS value increased as the ability difference between the two groups increased. The greater the difference in ability between the two groups and the smaller the sample size, the more pronounced the advantage of the 3E approaches. When the ability difference between the two groups was 0.1, only when the sample size was 200, the BIAS values of the PrWA_100 and PrWA_50 methods were smaller than that of the reference method; when the ability difference was 0.3, and the sample size was 200, 500 and 1000, the BIAS values of the PrWA_100 and PrWA_50 methods were smaller than that of the reference method. Whereas, when the ability difference was 0.5, the BIAS values of the PrWA_100

and PrWA_50 methods were smaller than that of the reference method, regardless of the sample size.

Among the five 3E approaches, the BIAS values tended to decrease as the powers of the precision used in the weighting schemas increased. Still, the values may not continue to fall as the power increased to 50, especially when the ability difference between the two groups was large.

The sample size has no uniform effect on the reference method, and for the 3E approaches, their BIAS values at a sample size of 2000 were always greater than those at a sample size of 200. However, this did not mean the BIAS value consistently increased according to the sample size. For example, the BIAS values of all the compared 3E approaches under the condition of 1000 sample size were all smaller than that for the 500-sample size condition.

In summary, in terms of equating accuracy, the higher the power of precision used in the weighting schemas for the 3E approaches, the better the performance. When the number of powers reaches 50, it is enough to outperform the eight reference learners in most cases. In addition, the greater the ability difference between the two groups and the smaller the sample size, the more noticeable the advantages of the 3E approaches (i.e., the PrWA_100 and PrWA_50 methods) over the reference method.

## Empirical Study

To showcase the performance of the new method in a practical setting, we present an empirical study using the final examination scores of fifth-year undergraduate students from a medical school. The surgery exam consisted of two rounds, corresponding to two tests, each containing 40 dichotomously scored items, with 12 common items between them. A total of 201 students were randomly assigned to the two tests. The descriptive information of raw scores can be found in Table 4.

**Table 4**

*Descriptive Statistics of the Total Scores for Each Group*

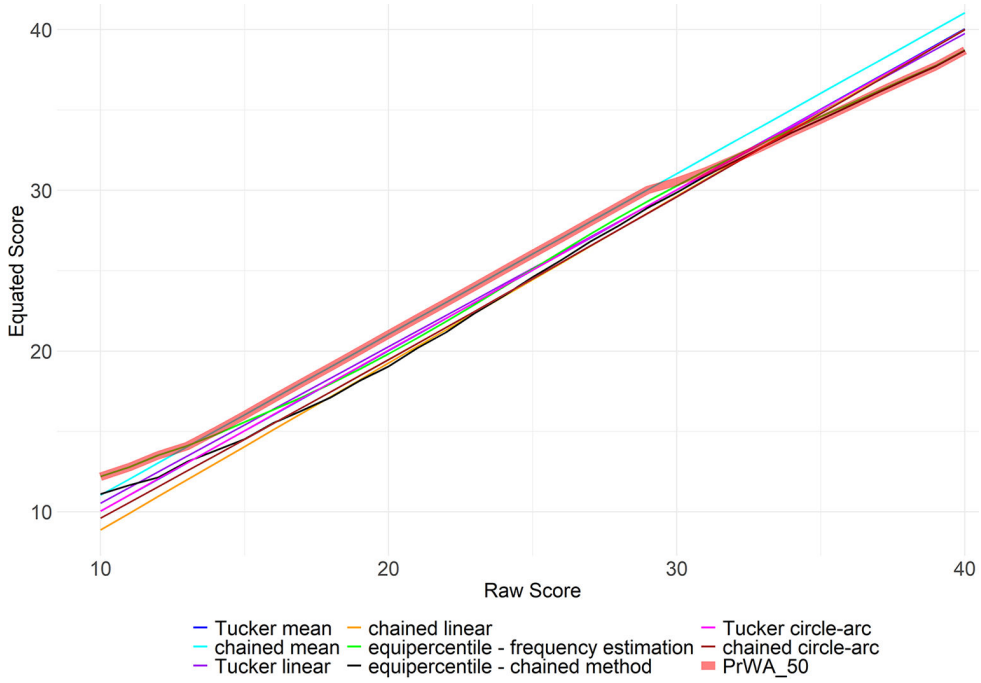
| | | Raw scores | | | |
|---|---|---|---|---|---|
| **Round** | ***N*** | **Minimum** | **Mean** | **Maximum** | **Average score on the anchor test** |
| 1 | 100 | 13.0 | 27.9 | 38.0 | 8.3 |
| 2 | 101 | 13.0 | 29.1 | 38.0 | 9.1 |

The eight equating methods used in the simulation study, along with the PrWA_50 method, were employed to equate the scores of the first round to those of the second round. The dataset and the code used can be obtained by contacting the corresponding

author. The equating results are displayed in Figure 4. Since no student's total score was below 10, the figure only shows the raw scores from 10 to 40. Although the true equating transformation is not known in the real dataset, it can be observed that the differences among various equating methods are not substantial. The PrWA_50 method is shown with a thicker red line, and in the score range of 10 to 30, it is closest to the results of the chained mean equating method. When the scores are above 30, it is more similar to the results of the two equipercentile equating methods. In other words, the PrWA_50 method can combine the results of various equating methods with different weights across different score ranges.

**Figure 4**

*Equating Results for a Real-World Dataset*




## Discussion and Conclusion

EL is a powerful technique that utilizes multiple models to improve the reliability and accuracy of individual models. This study proposed an empirical ensemble equating (3E) approach that treats multiple equating functions as learners in EL and adopted several weighting schemes to improve the equating accuracy under the NEAT Design. The

simulation study found that the 3E approach with weights proportional to the precision's powers of 50 or 100 can yield more accurate equating results than the eight ensembled equating methods in most cases. The 3E equating approach proposed in this study can better support the exchangeability of scores across different forms of an assessment, thereby guaranteeing the fairness of the assessment and providing support in constructing item banks more scientifically.

Holland and Strawderman (2009) introduced several approaches to averaging two or more equating functions, provided details on how to weigh the equating parameters, and discussed some properties of the averages of equating functions. The traditional methods of averaging equivalence functions introduced by Holland and Strawderman (2009) (e.g., the point-wise weighted average method, the angle bisector method, and the symmetric weighted average method) can be seen as following the ensemble idea but limited to linear equating functions or two nonlinear equivalence functions. The 3E approach proposed in this study can ensemble various linear or nonlinear equating methods and adopt various weighting schemes, which is more generalized and flexible, as the 3E eventually turns an equated score sheet in rather than a model.

Another perspective of understanding the utility of the proposed 3E approach is comparing it to the prior sensitivity analysis in Bayesian analysis. From the perspective of Bayesian analysis, the ideal priors should accurately reflect preexisting knowledge of the world, both in terms of the facts and the uncertainty about those facts. Priors that do not correspond to reality, however, can lead to severe bias (e.g., Baldwin & Fellingham, 2013; van Erp et al., 2019). Thus, a prior sensitivity analysis aiming to update one's prior beliefs with the data is vital to improve the performance of Bayesian modeling. Compared to typical prior sensitivity analysis, improving equating performance by proposing multiple equating designs is scarce in equating studies. There are commonalities between the proposed ensemble learning method with Bayesian prior sensitivity analysis. First, both utilize multiple methods (frequentist perspective) or priors' settings (Bayesian perspective) to improve prediction accuracy. The selection process could be arbitrary or purposeful for both frameworks depending on the research purpose. For instance, researchers may include the chained linear equating method as one learner because she/he, as an equating expert, believes the method is appropriate. Previous Bayesian literature has shown that some models perform better than the frequentist model when priors reflect researchers' beliefs appropriately (van Erp et al., 2019). Second, selecting those sources reflects different hypotheses (frequentist perspective) or beliefs (Bayesian perspective). In the simulation study, eight learners are chosen for learning. Like informative and uninformative prior, in ensemble learning, each learner reflects strong or weak hypotheses and thus performs better in specific scenarios than others. Third, some methods are needed to summarize the outputs of different outputs and obtain the best result. This study shows that EL is very flexible when adopting different average schemas. At the

same time, in Bayesian analysis, the prior setting could be updated using the information of previous prior settings to get better performance.

To develop further from this initial attempt, research directions can extend to setting comparison and monotonicity constraints. The former question relates to the assumptions for empirical settings of the 3E simulation scenarios; Are large-scale assessments more appropriate than the smaller ones as the estimated parameters are more stable and trackable due to their standardized nature? The second problem is that the equated scores do not grant a straight ascending/descending order; Can smoothing functions be used to ease the monotonicity violation? Thirdly, only the absolute bias has been used to calculate weights. It might be beneficial to consider both the equating bias and standard deviation when determining the weights for the ensemble approach. Finally, it's essential to remember that the 3E approach highly relies on the learners' qualities. Even though the contaminations by bad learners can be well controlled within the 3E approach's framework, a relatively large number of bad learners can still be detrimental to the final equated estimates.

**Funding:** This work was funded by the National Natural Science Foundation of China for Young Scholars (Grant No. 72104006), Peking University Health Science Center (Grant No. BMU2021YJ010), and the Peking University Health Science Center Medical Education Research Funding Project (Grant No. 2022YB41).

**Acknowledgments:** The authors have no additional (i.e., non-financial) support to report.

**Competing Interests:** The authors have declared that no competing interests exist.

**Data Availability:** The datasets analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request. Requests to access these datasets should be directed to YH, hanyuting@bjmu.edu.cn.

## Supplementary Materials

The supplementary materials provided is the R code (Version 4.2.2 64-bit), and can be accessed in the Index of Supplementary Materials below.

### Index of Supplementary Materials

Jiang, Z., Han, Y., Zhang, J., Xu, L., Shi, D., Liang, H., & Ouyang, J. (2023). *Supplementary materials to "Empirical ensemble equating under the neat design inspired by machine learning ideology"* [R code]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.12949

## References

Abidi, S. M. R., Zhang, W., Haidery, S. A., Rizvi, S. S., Riaz, R., Ding, H., & Kwon, S. J. (2020). Educational sustainability through big data assimilation to quantify academic procrastination using ensemble classifiers. *Sustainability, 12*(15), Article 6074. https://doi.org/10.3390/su12156074

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika, 82*(1), 48–66. https://doi.org/10.1007/s11336-016-9528-7

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18*(2), 151–164. https://doi.org/10.1037/a0030642

Borovkova, S., & Tsiamas, I. (2019). An ensemble of LSTM neural networks for high-frequency stock market classification. *Journal of Forecasting, 38*(6), 600–619. https://doi.org/10.1002/for.2585

Diao, H., & Keller, L. (2020). Investigating repeater effects on smallsample equating: Include or exclude? *Applied Measurement in Education, 33*(1), 54–66. https://doi.org/10.1080/08957347.2019.1674302

Holland, P. W., & Strawderman, W. E. (2009). How to average equating functions, if you must. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 89–107). Springer. https://link.springer.com/book/10.1007/978-0-387-98138-3

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement, 47*(3), 286–298. https://doi.org/10.1111/j.1745-3984.2010.00114.x

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking: Methods and practices* (3rd ed.). Springer.

Kumar, A., & Mayank, J. (2020). *Ensemble learning for AI developers.* BA Press.

Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2021). Targeting customers for profit: An ensemble learning framework to support marketing decision-making. *Information Sciences, 557*, 286–301. https://doi.org/10.1016/j.ins.2019.05.027

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

Moses, T., & Holland, P. W. (2010). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement, 47*(1), 76–91. https://doi.org/10.1111/j.1745-3984.2009.00100.x

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2019). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychological Medicine, 49*(14), 2330–2341. https://doi.org/10.1017/S003329171800315X

Premalatha, N., & Sujatha, S. (2021, September). An effective ensemble model to predict employment status of graduates in higher educational institutions. In *2021 Fourth International*

*Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICECCT52121.2021.9616952

Priore, P., Ponte, B., Puente, J., & Gómez, A. (2018). Learning-based scheduling of flexible manufacturing systems using ensemble methods. *Computers & Industrial Engineering, 126*, 282–291. https://doi.org/10.1016/j.cie.2018.09.034

Ragab, M., Abdel Aal, A. M., Jifri, A. O., & Omran, N. F. (2021). Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques. *Wireless Communications and Mobile Computing, 2021*, Article 6241676. https://doi.org/10.1155/2021/6241676

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*(3), 261–285. https://doi.org/10.1111/j.1745-3984.2010.00113.x

Stamp, M., Chandak, A., Wong, G., & Ye, A. (2021). On ensemble learning. In M. Stamp, M. Alazab, & A. Shalaginov (Eds.), *Malware analysis using artificial intelligence and deep learning* (pp. 223–246). Springer.

van der Linden, W. (2011). Local observed-score equating. In A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 201–223). Springer.

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology, 89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer.

Wagner, J. G. (1975). *Fundamentals of clinical pharmacokinetics*. Drug Intelligence Publications.

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. https://doi.org/10.1111/jedm.12212

Zeng, L. (1993). A numerical approach for computing standard errors of linear equating. *Applied Psychological Measurement, 17*(2), 177–186. https://doi.org/10.1177/014662169301700207