Evaluating the Accuracy of the Empirical Item Characteristic Curve Preequating Method in the Presence of Test Speededness

Educational and Psychological
Measurement

2019, Vol. 79(2) 288–309

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164418777854

journals.sagepub.com/home/epm

Yuxi Qiu¹ and Anne Corinne Huggins-Manley¹

Abstract

This study aimed to assess the accuracy of the empirical item characteristic curve (EICC) preequating method given the presence of test speededness. The simulation design of this study considered the proportion of speededness, speededness point, speededness rate, proportion of missing on speeded items, sample size, and test length. After crossing all of the manipulated factors and then normalizing the evaluation criteria (bias and root mean square difference [RMSD]) with regard to test length, the results revealed that (1) when test speededness was present, conversions from the EICC preequating method tended to be positively distorted; (2) no practically meaningful moderation effect associated with sample size was found on the relationship between test speededness and the accuracy of EICC preequating; and (3) the location of the speededness point was the driving factor in terms of its impact on the accuracy of EICC preequating. Implications and suggestions were discussed.

Keywords

EICC, item preequating, test speededness

---

¹University of Florida, Gainesville, FL, USA

Corresponding Author:

Yuxi Qiu, School of Human Development and Organizational Studies in Education, University of Florida, 1215 Norman Hall, Gainesville, FL 32611, USA.

Email: yqiu2013@ufl.edu

Introduction

For testing programs that continually develop new forms of a test, score equating is an essential approach to transform scores into a common scale, so that scores on different test forms have the same meaning over time (Dorans, Moses, & Eignor, 2009). Score equating typically takes place after a new test form is administered operationally, hence the name “postequating.” However, this practice is being challenged by the passing of disclosure legislation, which mandates that testing programs reveal test items contributing to an examinee’s reported test score after score reporting, which consequently raises issues in the realm of test production and security. For instance, Fremer (1981) reviewed the impacts of the New York State test disclosure legislation (i.e., the LaValle Act) on national testing programs such as the Graduate Management Admission Test (GMAT) and the Graduate Record Examination Aptitude Test (GRE) that were administered by Educational Testing Service (ETS). The review demonstrates that the legislation led to an increase in the number of test forms developed, accelerated the development of new equating methods, and increased filing requirements. He also discussed preequating (and section preequating in particular) as a possible solution to respond to the demands placed on testing agencies as a result of disclosure mandates.

According to Lord (1975), preequating can be defined operationally using the following scenario:

Suppose a pool of pretested items is available and that parameters have already been obtained for all items. It is now possible to build a series of tests and to place all the test scores on the same scale. The curvilinear scaling formula that affects this equating can be derived even before the test is printed or administered. (pp. 29-30)

Though Lord gave the above definition in the context of item response theory (IRT), it is versatile enough to highlight the power of preequating regardless of the psychometric framework for score calibration and equating. That is, conversions from raw to scaled scores can be established before a test form is administered intact, given that newly written items have been pretested and precalibrated (Kolen & Brennan, 2014). Other benefits accrue from preequating, beyond the ability to compare scores using postequating. For example, according to early research (e.g., Bejar & Wingersky, 1982; Eignor & Cook, 1983; Lord, 1980; Zu & Puhan, 2014), preequating affords (1) a reduction in the span of time required to report scores, (2) a remedy for testing agencies hoping to respond to the test disclosure legislations discussed above, (3) greater control of the exposure rates of new test forms, and (4) more time for quality control of established conversions, which are often completed in a hurried fashion during postequating.

To accrue these benefits, a psychometric framework must be selected to perform the preequating method. If item-level preequating is the goal, there are two previously studied approaches to the method: IRT-based preequating and empirical item characteristic curve (EICC) preequating. The former uses parameter estimates from

IRT models, whereas the latter uses a nonparametric approach. This study focuses on the nonparametric, EICC approach to preequating.

However, item preequating, either IRT or EICC based, can be vulnerable to the presence of context effects, test speededness effects, position effects, and multidimensionality (e.g., Bejar & Wingersky, 1982; Eignor & Stocking, 1986; Kolen & Harris, 1990; Livingston, 1985; Stocking & Eignor, 1986; Zu & Puhan, 2014). Test speededness comes about from the necessity of timing tests to maintain standardization, and it introduces an additional concern into item response calibration and equating. Zu and Puhan (2014) noted that the factor of speededness in particular can influence the accuracy of EICC preequating, but they did not study this concern directly. Given the lack of research on the influence of speediness on EICC preequating, and in addition the paucity of previously published research on EICC methods in general, this study hopes to increase knowledge of the potential benefits and challenges of preequating by assessing factors related to test speededness that may have non-negligible effects on the quality of EICC preequating.

The rest of this study discusses in further detail the existing research and procedures on IRT-based and EICC item preequating. Then, research questions are presented along with the corresponding simulation design. Finally, results from the simulation are discussed and recommendations are made.

A Further Review of Item Preequating

At the risk of oversimplifying, what distinguishes preequating from postequating is when the raw-to-scale conversions are produced during the course of testing. A flowchart (Figure 1) is provided to visualize the process of item preequating. In Figure 1, different forms of the same test are represented by rectangles with grayed-out areas and white areas corresponding to nonoperational and operational items, respectively, and an item pool is represented by a cylinder. Starting from the left-hand side of this flowchart, item preequating requires that nonoperational (i.e., unscored) items be pre-administered with operational (i.e., scored) ones, and that analytical techniques such as IRT and EICC then be used to calibrate items that are both operational and nonoperational. Note that it is necessary to translate items from different test forms into the same scale. So far, a pool of items (represented by the cylinder in Figure 1) has been created. In the development of future test forms, shown in the right-hand side of the flowchart, a new form can be assembled with the operational portion consisting of items from the existing item pool and the nonoperational portion comprising newly written items. Once a new form is constructed, item preequating can be implemented across the operational portion of the form to establish the raw-to-scale conversions.

Procedures of Item Preequating: IRT and EICC

Calibration. IRT methods use a finite set of parameters (e.g., latent ability— θ , item discrimination— a , item difficulty/location— b , and lower asymptote— c) to define the probability that an examinee will select a particular test response. To the extent

Figure 1. Procedures of item preequating (summarized based on Kolen & Brennan, 2014).

that the assumptions of unidimensionality and local independence are met, IRT methods afford considerable flexibility in areas such as test development, adaptive testing, equating, and so forth (Kolen & Brennan, 2014). Estimations of item parameters and latent ability are typically achieved using joint maximum likelihood (JML), marginal maximum likelihood (MML), conditional maximum likelihood (CML), or Bayesian methods (Yen & Fitzpatrick, 2006).

The EICC method, on the other hand, is data driven. As Livingston (1984) described, EICCs are formed by grouping examinees according to their scores on a criterion variable and then obtaining the proportion of correct responses for each item at each level of the chosen criterion. Unlike IRT methods, the EICC method does not use the aforementioned estimation techniques but instead relies on smoothing techniques to handle irregularities in the relationships between proportions of correct responses and the chosen criterion. Examples of these different procedures employed for the EICC method include the polynomial function in Livingston (1984) and the Gaussian kernel smoothing in Zu and Puhan (2014). In addition, the quality of calibration in the EICC method is affected by the selection of kernel smoothers and bandwidth (or the intensity of the degree in the polynomial function).

Scale Transformation. As discussed at the beginning of this section, a procedure of scale transformation must be implemented with items from different forms of the same test. This procedure assures that items in the same pool are comparable in terms of their (E)ICCs. In IRT methods, transformations of scale are accomplished through either concurrent calibration or separate calibration with mean/mean, mean/sigma, or characteristic curve methods (e.g., the Haebara approach and the Stocking-Lord approach; Kolen & Brennan, 2014). By contrast, in the EICC method, according to Zu and Puhan (2014), methods based on observed scores can be used to equate

scores on the criterion variable. In their research, Zu and Puhan employed equipercentile equating to transform scales.

Preequating. For a new form of a test, items in the operational portion can be preequated prior to the administration of that particular form. What makes IRT and EICC ideal for item preequating is the invariance nature of (E)ICCs. Specifically, in IRT methods, item parameters are invariant across test forms, and likewise EICCs are assumed to remain the same in new test forms (Zu & Puhan, 2014). In IRT-based item preequating, ICCs are summed across items in the operational portion to obtain the test characteristic curve (TCC) for the new test form. Then, following the true-score equating process outlined by Kolen and Brennan (2014), the equated score of an expected score on the new form would be the expected score on the old form that corresponds to the same latent ability— θ associated with the expected score on the new form. A Newton–Raphson method can be used in the foregoing process.

According to Livingston (1984), the logic of EICC item preequating is akin to that of the IRT-based approach, except that the empirical TCC is conditioned on the scores of the criterion variable, not the latent trait. Then, concordances between the new form and the old form can be established by identifying expected scores on both forms that correspond to the same score on the chosen criterion. In Livingston's (1984) study, regression was used to determine the concordance.

Previous Research on Item Preequating: IRT and EICC

IRT Preequating. The origins of item preequating can be traced back to Marco (1977), whose study compared raw-to-scale conversions established from IRT preequating to those from IRT postequating on the SAT Verbal test sections administered from 1968 to 1974. The results indicated that IRT preequating tended to introduce positive bias for the full-scaled score range of 200 to 800 but presented the least amount of error in the range of 400 to 600 where college admissions decisions were made. A number of studies investigating the performance of IRT preequating methods were also conducted in late 1990s. For instance, Bejar and Wingersky (1982) assessed the feasibility of IRT preequating based on the Test of Standard Written English (TSWE); Eignor and colleagues conducted a series of studies on IRT preequating based on data collected from the SAT (Eignor, 1985; Eignor & Cook, 1983; Eignor & Stocking, 1986; Stocking & Eignor, 1986); and Kolen and Harris (1990) experimented with item preequating in the American College Test's (ACT) Mathematics Assessment Test. The most recent research on the feasibility of IRT preequating in the context of large-scale testing programs is seen in Domaleski (2006); Tong, Wu, and Xu (2008); and Gao, He, and Ruan (2012). The studies show that IRT-based preequating achieves mixed results. Potential factors contributing to its uneven performance include context effects, test speededness, item position effects, and multidimensionality. Even so, IRT-based item preequating is still in demand in large-scale testing programs (Zu & Puhan, 2014; Gao et al., 2012).

EICC Preequating. As discussed previously, the flexibility of IRT is based on strong assumptions that may be difficult to meet in practice (Kolen & Brennan, 2014). For example, in their research on a collection of tests with purposes varying from measuring basic skills to assessing college readiness, and test forms varying from traditional paper-and-pencil to computer-based, Sinharay, Haberman, and Jia (2011) reported evidence that results obtained from IRT models did not adequately fit with the data of the given sets of tests.

Use of EICC item preequating was first seen in Livingston (1984, 1985) to preequate the four forms (two Verbal and two Math) of the New Jersey College Basic Skills Placement Test. Adequacy of EICC preequating was found on three out of the four test forms. Recently, Zu and Puhan (2014) evaluated this method using data from a certification testing program. Based on this dataset, they created three parallel test forms (X_1 , X_2 , and X_3 —the new form) and made two subgroups of examinees take X_1 and X_2 . They then compared results from EICC and three-parameter logistic (3-PL) IRT item preequating to the results from a single group equipercentile equating. They found that EICC preequating showed smaller deviances from the equipercentile equating compared with 3-PL IRT preequating. In addition, both preequating methods deviated from the equipercentile equating at the lower and higher tails of the frequency distribution of the new form. It is worth mentioning that they used the total scaled score as the criterion variable and Gaussian kernel smoothing to handle irregularity rather than employing an anchor-test score and polynomial smoothing, as in Livingston (1984). In addition, 3-PL IRT preequating was included for the purpose of comparing the results to EICC preequating, despite the fact that the chosen IRT model did not fit the data adequately. Study results such as those summarized here, along with the relative paucity of research on EICC preequating, indicate the need for additional research on EICC preequating and the conditions under which it may or may not perform well.

Purpose

The purpose of this study was to evaluate the performance of EICC preequating under test speededness situations that are commonly observed in operational test administrations. Because test speededness can come in many forms and the effects of speededness may depend on several factors, the study varied multiple facets related to speededness, including when speededness begins (speededness point); how quickly the probability of correct response decays due to speededness (speededness rate); how many items are omitted due to speededness (proportion of missing); and how many examinees are impacted by speededness (proportion of persons experiencing speededness). In order to systematically investigate the impact of the previously mentioned factors related to speededness, a simulation study was conducted and the following questions were investigated:

1. To what extent do factors related to test speededness affect the accuracy of the results of EICC preequating?
2. Does sample size moderate the relationship between speededness and the accuracy of EICC preequating?

Method

Simulation Design

Data were simulated based on the 3-PL version of the speededness-generating model (Wollack & Cohen, 2004), which is defined as

$$P_{ij}^* = c_i + (1 - c_i) \left\{ P_i(\theta_j) \cdot \min \left(1, \left[1 - \left(\frac{i}{n} - \eta_j \right) \right]^{\lambda_j} \right) \right\} \quad (1)$$

where $P_i(\theta_j)$ is the probability of correctly responding to item i for examinee j under the 2-PL IRT model, and $\eta_j (0 \leq \eta_j \leq 1)$ and $\lambda_j (\lambda_j \geq 0)$ denote the speededness point parameter and the speededness rate parameter for examinee j , respectively. According to Wollack and Cohen (2004), η_j represents the proportion of items that examinee j , proceeding from the beginning of the test, will complete before experiencing speededness. Hence, a smaller value of η_j indicates that examinee j tends to experience speededness earlier in the test. Before the speededness point $\left(\frac{i}{n} \leq \eta_j \right)$, the function $\min(\cdot)$ returns a value of 1, and the above formula reduces to a 3-PL IRT model. Once an examinee passes the speededness point $\left(\frac{i}{n} > \eta_j \right)$, $\min(\cdot)$ returns $1 - \left(\frac{i}{n} - \eta_j \right)$, which in turn is raised to the power of λ_j that controls the rate of decrease of $P_i(\theta_j)$. When $\eta_j = 1$ or $\lambda_j = 0$ the model reduces to a 3-PL IRT model. For example, $\eta_j = .5$ for a 10-item test ($n = 10$) indicates that examinee j would begin to experience speededness after completing half of the items. Hence, when examinee j does not pass the speededness point ($i \leq 5$), $\min(\cdot)$ returns a value of 1 and equation (1) reduces to a 3-PL IRT model regardless the value of λ_j . When examinee j passes the speededness point and is answering the sixth item—where ($i = 6$), then $\min(\cdot) = .9$ —the probability of correct response of examinee j will be adjusted downward for $\lambda_j > 0$.

This speededness-generating model has been selected as the data generation model to reflect the nature of speededness as a complex phenomenon of testing, and there is no reason to assume all examinees follow the same pattern in terms of their interaction with speededness. For a given test, as Wollack and Cohen (2004) suggested, examinees can experience speededness at different points and take various strategies to cope with time pressure. The chosen model is appealing because it takes into account both the variability of the point when examinees begin to experience speededness (η) and its impact on the probability of correct response (λ).

Table 1. Preequating Design.

Test form (30/60/90)	Item bank ^a (45/90/135)			Pool of examinees (4000/2000/1000)	
	A	B	C	P (2000/1000/500)	Q (2000/1000/500)
Y	A	B		P	—
X ₁	A		C	—	Q
X ₂ ^b		B	C	—	—

^aA, B, and C are subsets of items randomly selected from the simulated item bank and contain an equal number of items.

^bX₂ is the new test built from nonoperational components of Form Y and Form X₁; Form X₂ is preequated to Form Y.

Item parameters were randomly sampled from the following distributions that mimic a test with a wide range of discrimination, difficulty, and lower asymptote parameters. Specifically, $a \sim \mathcal{U}(0.55, 1.36)$, $\ln(b) \sim \mathcal{N}(0, 1)$, and $c \sim \mathcal{U}(0, 0.3)$. Ability parameters were generated from a normal distribution $\theta \sim \mathcal{N}(0, 1)$. For examinees in a speeded class (i.e., a subset of examinees who experienced some form of speededness on some test items), speededness point and speededness rate parameters were generated from a beta distribution and a lognormal distribution, respectively (Suh, Kang, Wollack, & Kim, 2006). Manipulations of both parameters were discussed in subsequent subsections.

In order to maximize the generalizability of the findings of this research, the following factors were manipulated to emulate previous studies.

Sample Size (3-Level). The number of examinees taking each test form was varied at three levels: 2,000, 1,000, and 500. A value of 2,000 was selected to mimic the Livingston (1984) study, in which each new item was pretested on more than 2,000 students. Reductions of sample size were intended to stress this preequating method. These within-form sample sizes aligned with total examinee pool sample sizes of 4,000, 2,000, and 1,000 (Table 1).

Test Length (3-Level). Test length was manipulated at three levels in terms of the number of items: 30, 60, and 90, which are all commonly observed in previous research (e.g., Guo & Puhan, 2014; Oshima, 1994). Item banks of size 45, 90, and 135 were generated at the beginning, and then equivalent test forms with intended length were developed accordingly. As discussed below, we opted to normalize the evaluation criteria according to this test length condition, so this factor is not directly evaluated for its influence on the results. Rather, this factor is included to ensure that the results we present generalize across various test lengths.

Location of Speededness Point (3-Level). The term “location” in this case represents the first item on a given test at which an examinee begins to experience speededness.

Manipulation of this factor was accomplished by varying the mean of the generating distribution of η . Specifically, examinees in speeded class η were generated from either Beta(120, 80), Beta(240, 80), or Beta(750, 80). Correspondingly, on average, 40%, 25%, and 10% of the items from the end of a test were considered as showing speededness. The conditions with 40% of speeded items mirrored a study by Suh et al. (2006), and we added the conditions of 25% and 10% to correspond with more moderate and slightly speeded conditions. η was fixed at 1 for examinees in the non-speeded class, which represents a situation in which examinees completed all test items without experiencing speededness.

Magnitude of Speededness Rate (3-Level). The average rate of decay (i.e., λ) in terms of the probability of correct response on an item was manipulated by setting the geometric mean of the log normal distribution at three levels: 3.912, 2, and 0.5. The average speededness rate of 3.912 was chosen to resemble the condition in the study by Suh et al. (2006). The conditions of 2 and 0.5 were added to correspond with lower and more moderate rates of decay. λ was fixed at 0 for examinees in the nonspeeded class.

Proportion of Speededness (2-Level). This represents the proportion of examinees experiencing speededness, where two levels were simulated to correspond to low (10%) and high (30%). For the low-speededness condition, 10% of the total examinees experienced speededness. On the other hand, 30% of the total examinees experienced speededness for the high-speededness condition. These two levels were chosen to resemble the study by Suh, Cho, and Wollack (2012).

Proportion of Missing (2-Level). Among the items showing speededness, two different percentages of missing responses were simulated: 20% and 40%. According to Suh et al. (2012), these values are realistic for actual testing programs.

Summary of Simulation Design. By completely crossing the above factors, a total of 324 unique conditions were formed. In addition to these conditions with the presence of test speededness effects, a set of baseline conditions in which no examinees experienced speededness was introduced into each combination of sample size and test length. Throughout this study, codes for data generation and calibration of EICCs—including those performing Gaussian kernel smoothing—were written in base R (R Core Team, 2017). In addition, R-package “dplyr” (Wickham, Francois, Henry, & Müller, 2017), “ggplot2” (Wickham, 2009), and “equate” (Albano, 2016) were used for data manipulation, data visualization, and scale transformation (discussed in the subsequent subsections), respectively. Thirty replications were run for each condition.

Preequating Design

The preequating design followed the Zu and Puhan (2014) study. Specifically, a pool of items was generated at the beginning, and then items were equally divided into

three component parts named A, B, and C. After that, three test forms were created by combining two of the above components (Table 1). Form Y and X_1 were old forms containing speededness, and Form X_2 was the new test form that was to be preequated to the scale of Form Y. The test forms were considered randomly parallel, because components A, B, and C were created from items generated from the same distributions of item parameters (Hwang & Cleary, 1986). Next, a pool of examinees was generated and then split into half to create two samples, with sample P taking Form Y and sample Q taking Form X_1 .

Procedure of EICC Preequating

Following the procedures of item preequating discussed previously and the design in Zu and Puhan (2014), the EICCs for items in Form Y were estimated with the raw score scale of Form Y as the criterion variable, and a Gaussian kernel smoother was used to deal with irregularity. A bandwidth of $h = N^{-0.2}$ was selected according to the recommendation in Ramsay (1991), where N denoted the sample size. These items, with their EICCs, were considered as part of a newly developed item pool.

After the administration of Form X_1 to sample Q, scores on Form X_1 were transformed to the scale of Form Y. The intended scale transformation was completed with a chained equipercentile equating, given that half of the items in both test forms were in common (Component A). Then, the raw scores on Form X_1 were replaced with their Form Y equivalents, and EICCs for items in Form X_1 were estimated by using the equated scores as the criterion variable. It should be noted that EICCs for items in component C of Form X_1 were added to the item pool only.

From the established item pool, a new test form, X_2 , was created consisting of items in Components B and C (Table 1). Then, X_2 was preequated to Form Y. At the lower end of expected total scores, a score of 0 on X_2 was set equal to the score of 0 on Y, and the lowest expected total scores on X_2 were set equal to the lowest expected total on Y. Then, linear interpolation was used to convert scores outside the range of possible expected total scores. Likewise, similar steps were implemented to convert scores of X_2 outside the range of possible expected total scores, except that the highest expected total scores on X_2 and Y, as well as the maximum possible scores on X_2 and Y, were set equal, respectively. A similar procedure was also used in Zu and Puhan (2014).

Evaluation Criteria

To summarize the differences of equated scores of Form X_2 between speeded and nonspeeded testing contexts, two indices, bias and root mean square difference (*RMSD*), were used as measures of the accuracy of EICC preequating. Bias and *RMSD* were commonly used to summarize discrepancies of equated scores in previous research (e.g., Kolen & Harris, 1990). Low values in bias and *RMSD* suggest

fewer discrepancies between equated scores and consequently are indicative of greater accuracy. They can be calculated as

$$\text{bias} = \frac{\sum f_i (X'_i - X_i)}{\sum f_i} \quad (2)$$

and

$$\text{RMSD} = \sqrt{\frac{\sum f_i (X'_i - X_i)^2}{\sum f_i}} \quad (3)$$

where f_i refers to the frequency of integer score i on Form X_2 when there is no speededness effect; X is a score from Form X_2 that has been equated to the scale of Form Y ; X_i and X'_i differ only in that the X_i scores came from conditions of no speededness and X'_i scores came from conditions of speededness. In order to facilitate fair comparisons across conditions, bias and RMSD were normalized based on the number of items.

Results

Visualization of Bias and RMSD

Figures 2-5 visualize the change of average bias relative to the four facets of speededness, which include the proportion of examinees experiencing speededness during testing (Figure 2), the location where examinees begin experiencing speededness (Figure 3), the average rate of decay in terms of the probability of correct response on speeded items (Figure 4), and the proportion of missing on speeded items (Figure 5).

Proportion of Speededness. While raising the proportion of examinees experiencing test speededness from 10% to 30%, a general inflation of bias was observed in Figure 2. This is indicative of a greater discrepancy of equated scores from EICC preequating between conditions with and without speededness, when approximately 30% examinees experienced speededness. However, the inflation appears to vary when the other five factors (speededness point, speededness rate, proportion of missing, test length, and sample size) change at different levels. Holding speededness rate, proportion of missing, test length, and sample size constant, for instance, the increase of bias becomes less apparent when the speededness point moves toward the end of the test.

Speededness Point. From Figure 3, the magnitude of bias appears to decline as the speededness point changes from 60% to 90%. As described in the methods section, a value of 60%, 75%, and 90% of the speededness point represents the situation in which examinees are able to complete 60%, 75%, and 90% of items for a given test without speededness. Such a situation then corresponds to 40%, 25%, and 10% of items characterized by speededness for this particular test, respectively. The observed

Figure 2. Change of bias over the levels of proportion of speededness. Each grid represents a unique combination of the other five factors. Values on the top correspond to speededness point (60%—Beta(120, 80), 75%—Beta(240, 80), 90%—Beta(750, 80)), speededness rate (0.05, 2, 3.912), and proportion of missing (20%, 40%), respectively. Values on the right represent sample size (2,000, 1,000, 500) and test length (90, 60, 30).

Figure 3. Change of bias over the levels of speededness point. Each grid represents a unique combination of the other five factors. Values on the top correspond to test length, speededness rate, and proportion of missing, respectively. Values on the right represent sample size and proportion of speededness.

Figure 4. Change of bias over the levels of speededness rate. Each grid represents a unique combination of the other five factors. Values on the top correspond to test length, speededness point, and proportion of missing, respectively. Values on the right represent sample size and proportion of speededness.

Figure 5. Change of bias over the levels of proportion of missing on speeded items. Each grid represents a unique combination of the other five factors. Values on the top correspond to test length, speededness point, and speededness rate, respectively. Values on the right represent sample size and proportion of speededness.

decline of bias suggests that deviations of equated scores for conditions with speededness from conditions without speededness become smaller when the completion rate of items without speededness rises. In addition, the magnitude of decline seems drastic, which would indicate a strong effect of the speededness point on the accuracy of EICC preequating.

Speededness Rate. As mentioned previously, the speededness rate controls the average rate of decay of one's probability of correct response. The greater the value, the faster it declines. As shown in Figure 4, the changes of bias along the speededness rate are somehow mixed in terms of their directions. Specifically, when the speededness point has the values of 75% or 90% completion for a given test length, bias moves upward slightly or appears unchanged as the speededness rate rises from 0.5 to 3.912. But when examinees start experiencing speededness after completing 60% of items (i.e., speededness point = 60%), there is a general decrease in bias for the condition of 40% missing as the speededness rate increases and a general increase in bias for the condition of 20% missing. This can be indicative of an interaction between speededness and the other factors, and it requires a further empirical analysis (refer to the analysis of variance [ANOVA] in subsequent section) to uncover their relationships.

Proportion of Missing. According to Figure 5, the proportion of missing on speeded items presents a positive effect on the magnitude of bias. That is, as the former moves upward, bias tends to increase. There is also evidence of an interaction between the proportion of missing and the other factors. For example, the impact of the proportion of missing becomes less apparent for a speededness point of 90%.

Changes in bias for the aforementioned four factors were consistent in the sense that increasing the proportion of speededness, shifting backward the location of speededness point, accelerating the speededness rate, and increasing the proportion of missing all appear to be associated with an increase in bias (i.e., a loss of accuracy of EICC preequating).

Similar patterns of change were observed for *RMSD* as well, but the corresponding visualizations are omitted from this text because of space constraints. Plots of change of *RMSD* are available on request from the authors.

Analysis of Main and Interaction Effects on Bias and RMSD

From the above results of bias and *RMSD*, it is clear that factors related to test speededness have some degree of impact on the accuracy of EICC preequating. Hence, a factorial ANOVA was performed to further explore the extent to which these factors, and potential interactions therein, affect EICC preequating accuracy.

The assumptions of normality and homogeneity of variance were tested via examination of residuals and Levene's test, respectively. Review of *z*-scores of skewness and kurtosis for bias ($z_{\text{skewness}_{\text{bias}}} = \frac{.13}{.03} = 5.32$, $p < .05$; $z_{\text{kurtosis}_{\text{bias}}} = \frac{.86}{.05} = 17.2$, $p < .05$)

Table 2. Results of Mean Differences of Bias From Factorial Analysis of Variance (ANOVA).

Source ^a	df	MS	F	ω^2 ^b
Speededness point	2	1.20	39410.71***	.47
Proportion of speededness	1	0.96	31370.54***	.19
Proportion of missing on speeded items	1	0.41	13474.95***	.08
Sample size	2	0.13	4299.40***	.05
Proportion of speededness $\times$ Speededness point	2	0.12	4004.18***	.05
Proportion of missing on speeded items $\times$ speededness point	2	0.06	1845.19***	.02
Sample size $\times$ test length	4	0.02	735.785***	.02
Sample size $\times$ speededness point	4	0.01	406.64***	.01
Proportion of speededness $\times$ speededness rate	2	0.02	554.29***	.01

^aA factorial ANOVA model was used to compare mean differences of bias on the six factors in the simulation design. However, only the effects that are statistically significant and practically meaningful were reported.

^b $\omega^2 = .01, .06$, and $.14$ corresponds to small, medium, and large effect size, respectively (Cohen, 1988).
*** $p < .001$.

and $RMSD$ ($z_{skewness_{RMSD}} = \frac{.16}{.03} = 6.40, p < .05$; $z_{kurtosis_{RMSD}} = \frac{1.13}{.05} = 22.52, p < .05$) was indicative of problems with skewness and kurtosis. However, as recommended by Field (2013), due to a large sample size ($N = 9720$ in this study), a slight deviation from normality can be highly significant, and ANOVA seems robust to violate the assumption of normality, given a balanced design and the central limit theorem. On the other hand, Levene’s test of equality of error variances indicated that the assumption of homogeneity was not met for either bias ($F(323, 9396) = 3.48, p < .001$) or $RMSD$ ($F(323, 9396) = 4.26, p < .001$). However, heterogeneity was not a concern, because the research design was balanced with a sufficient sample size (Lomax & Hahs-Vaughn, 2012).

Tables 2 and 3 present part of the ANOVA results. Effects were selected to be reported based on the two criteria, specifically that they were statistically significant ($p < .05$) and practically meaningful. The latter was judged according to the Cohen’s (1988) rule of thumb: $\omega^2 = .01, .06$, and $.14$, corresponding to small, medium, and large effect sizes, respectively. We adopted a stance that any effects below Cohen’s (1988) definition of medium (i.e., $\omega^2 < .06$) would be considered negligible in practice, but that all other effects (i.e., $\omega^2 \geq .06$) could be practically meaningful.

For bias as the outcome (Table 2), the main effect of speededness point was the largest statistically significant effect. As the speededness point increases (i.e., the speededness begins later in the test), the bias decreases. The main effects of proportion of speededness and proportion of missing on speeded items were also significant and considered to be potentially practically meaningful, although the size of these effects was smaller than that of speededness point. For proportion of speededness, the lower proportion tested (i.e., 10%) resulted in smaller bias as compared with the higher proportion tested (i.e., 30%). For proportion of missing on speeded items,

Table 3. Results of Mean Differences of *RMSD* From Factorial Analysis of Variance (ANOVA).

Source ^a	<i>df</i>	<i>MS</i>	<i>F</i>	ω^2 ^b
Speededness point	2	1.57	51668.71***	.51
Proportion of speededness	1	1.07	35070.48***	.17
Proportion of missing on speeded items	1	0.49	15922.33***	.08
Sample size	2	0.13	4204.56***	.04
Proportion of speededness $\times$ Speededness point	2	0.15	5028.55***	.05
Proportion of missing on speeded items $\times$ speededness point	2	0.07	2370.27***	.02
Speededness rate	2	0.04	1258.18***	.01
Speededness point $\times$ sample size	4	0.01	451.66***	.01
Proportion of speededness $\times$ speededness rate	2	0.03	860.75***	.01

^aA factorial ANOVA model was used to compare mean differences of *RMSD* on the six factors in the simulation design. However, only the effects that are statistically significant and practically meaningful were reported.

^b $\omega^2 = .01, .06$, and $.14$ corresponds to small, medium, and large effect size, respectively (Cohen, 1988).

*** $p < .001$.

20% missing was associated with lower bias as compared with 40% missing. When *RMSD* was used as the outcome of the ANOVA models, results were very similar to those associated with the bias outcome (see Table 3).

Discussion

Test preequating, particularly IRT-based, has been widely used by testing programs as a method to cope with the test disclosure legislation, to bolster test security, and to reduce score reporting time (Bejar & Wingersky, 1982; Eignor & Cook, 1983; Lord, 1980; Zu & Puhane, 2014). However, application of this technique comes with its own challenges. For instance, context effects, test speededness, and other factors can impact the accuracy of conversions established from preequating. In addition, there is not much literature on this subject, which poses further challenges to practitioners and researchers in understanding the nature of various methods of preequating and how they interact with the previously mentioned factors that potentially affect the results of preequating. Furthermore, meeting the assumptions of IRT introduces additional concerns about the applicability of preequating. In this study, we assessed the performance of one item preequating method based on observed scores—namely, EICC preequating—as an alternative to IRT-based item preequating under various conditions related to test speededness.

The findings suggested that the results of EICC preequating tended to be positively biased in the presence of test speededness, meaning that equated scores were higher than they would have been under testing conditions without speededness. This, consequently, was indicative of a violation of the assumption of invariance of EICCs in the

presence of test speededness, which is consistent with what Zu and Puhan (2014) speculated but did not directly study.

After normalizing the evaluation criteria with respect to test length and then crossing all of the factors manipulated in this study, the main effect of speededness point presented a very large effect size on both bias and *RMSD* outcomes. Also, the main effect of proportion of examinees experiencing speededness and the main effect of proportion of missing data due to speededness were both associated with practically meaningful (i.e., moderate to large) effects on bias and *RMSD*. The interaction terms were all deemed as practically negligible. It appears that speededness point was the driving factor of increased bias and *RMSD* (i.e., degraded accuracy of EICC preequating), which indicates that EICCs are sensitive to and can lack invariance between administrations that vary in their speededness points. Therefore, a location of the speededness point that begins earlier in a test tends to result in more positive bias in EICC preequated scores than speededness that begins later in a test, and this effect is consistent over other testing conditions. The other two main effects located in this study are also expected to have practical significance in testing. As the proportion of examinees experiencing speededness increases, bias and *RMSD* are expected to increase. Also, as the proportion of missing data due to speededness increases, bias and *RMSD* are expected to increase.

In practice, when a test administration is characterized by test speededness, the item at which an examinee begins experiencing speededness would depend on his or her proficiency level and the order of items. Oshima (1994) has noted that for test developers who have control over the order of item presentation in a test, ordering items from easy to hard would minimize the impact of test speededness on item calibration and test equating. Our study supports this finding, even though we did not directly test item ordering effects. We also want to emphasize with our study the fact that test speededness can be a threat to test validity. Specifically, speededness can introduce bias into conversions established from item preequating, which consequently would distort interpretations and inferences drawn from preequated scores. Hence, test speededness should not be ignored and needs to be evaluated and controlled for when testing programs rely on precalibration for form development and on preequating for score reporting. When the effects of speededness are evaluated, we recommend focusing not only on the speededness point but also on the proportion of persons who were influenced by speededness and the proportion of missing data suspected to have arisen due to speededness. Practitioners may consider fitting models that can estimate parameters of speededness rate and speededness point that vary by examinees (e.g., see Equation 1), and use that information, as well as the observed proportion of missing responses for each examinee, to inform the expected accuracy of EICC preequating results.

There were several limitations to this research. In the current preequating design, the anchor tests were set at 50% of items in common between the old forms Y and X₁, which seems less likely to occur in practice. Thus, future research could be conducted to investigate how varying anchor test length influences the results from

EICC preequating, and how anchor test length interacts with factors related test speededness. Another limitation was the procedure of EICC preequating, in which only Gaussian kernel smoothing was considered. As Zu and Puhan (2014) mentioned, there are other smoothing techniques that can be useful. Researchers are encouraged to assess the performance of EICC preequating under different smoothing techniques. Finally, as discussed in the literature review, although preequating has been adopted by many testing programs to satisfy the needs of test security and the demand on short score reporting time, research on preequating is still in its infancy. Systematic comparisons between IRT-based and non-IRT-based item preequating as well as comparisons between item preequating and section preequating would be helpful to elucidate the characteristics of existing methods of preequating.

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) received no financial support for the research, authorship, and/or publication of this article.

References

- Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, 74(8), 1-36. doi:10.18637/jss.v074.i08
- Bejar, I. I., & Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement*, 6, 309-325.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Revised ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Domaleski, C. (2006). *Exploring the efficacy of preequating a large scale criterion-referenced assessment with respect to measurement equivalence* (Unpublished doctoral dissertation). Georgia State University, Atlanta.
- Dorans, N. J., Moses, T. P., & Eignor, D. R. (2009). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21-42). New York, NY: Springer.
- Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT Verbal and Mathematical sections* (Research Report 85-10). Princeton, NJ: Educational Testing Service.
- Eignor, D. R., & Cook, L. L. (1983). An investigation of the feasibility of using item response theory in the pre-equating of aptitude tests. Retrieved from <https://files.eric.ed.gov/fulltext/ED235197.pdf>
- Eignor, D. R., & Stocking, M. L. (1986). *An investigation of the possible causes for the inadequacy of IRT preequating* (Research Report 86-14). Princeton, NJ: Educational Testing Service.
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Thousand Oaks, CA: Sage.

- Fremer, J. (1981, April). *Impact of test disclosure legislation on test development*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.
- Gao, R., He, W., & Ruan, C. (2012). Does preequating work? An investigation into a preequated testlet-based college placement exam using postadministration data. *ETS Research Report Series*, 2012(1), i-18.
- Guo, H., & Puhan, G. (2014). Section preequating under the equivalent groups design without IRT. *Journal of Educational Measurement*, 51, 301-317. doi:10.1111/jedm.12049
- Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *An introduction to statistical concepts*. New York: Routledge.
- Hwang, C., & Cleary, T. A. (1986, April). *Comparing IRT and section pre-equating: A simulation study*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.
- Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, 27, 27-39.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer Science & Business Media.
- Livingston, S. A. (1984, April). *Item selection and pre-equating with empirical item characteristic curves*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.
- Livingston, S. A. (1985, April). *Large-sample pre-equating: How accurate?* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.
- Lord, F. M. (1975). A survey of equating methods based on item characteristic curve theory. *ETS Research Report Series*, 1975(1), i-35.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.
- Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31, 200-219.
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611-630.
- Sinharay, S., Haberman, S., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests* (Research Report 11-29). Princeton, NJ: Educational Testing Service.
- Stocking, M. L., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT preequating* (Research Report 86-49). Princeton, NJ: Educational Testing Service.
- Suh, Y., Cho, S. J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement*, 49, 285-311.
- Suh, Y., Kang, T., Wollack, J. A., & Kim, S. (2006, April). *A comparison of test scoring methods in the presence of test speededness*. Paper presented at the annual meeting of National Council on Measurement in Education Conference, San Francisco, CA.

- Tong, Y., Wu, S. S., & Xu, M. (2008, March). *A comparison of pre-equating and post-equating using large-scale assessment data*. Paper presented at the annual meeting of the American Educational and Research Association, New York, NY.
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.
- Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). dplyr: A grammar of data manipulation. R package version 0.7.4. Retrieved from <https://CRAN.R-project.org/package=dplyr>
- Wollack, J. A., & Cohen, A. S. (2004, April). *A model for simulating speeded test data*. Paper presented at the meeting of the American Educational Research Association. San Diego, CA.
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (pp. 111-153). Portsmouth, NH: Greenwood.
- Zu, J., & Puhan, G. (2014). Preequating with empirical item characteristic curves: An observed-score preequating method. *Journal of Educational Measurement*, 51, 281-300. doi:10.1111/jedm.12047