ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 455-471

The Impact of Item Preknowledge on Scaling and Equating: Item Response Theory True and Observed Score Equating Methods

Çiğdem AKIN ARIKAN* Allan S. COHEN**

Abstract

Testing programs often reuse items due mainly to the difficulty and expense of creating new items. This poses potential problems to test item security if some or all test-takers have knowledge of the items prior to taking the test. In this study, simulated data are used to assess the effect of preknowledge on item response theory true and observed score equating. Root mean square error and bias were used for the recovery of equated scores and linking coefficients for scaling methods. The results of this study indicated that item preknowledge has a large effect on equated scores and linking coefficients. Furthermore, as the mean ability distribution of the group difference, the number of exposed items, and the number of examinees with item preknowledge increase, the bias and RMSE for equated scores and linking coefficients also increase. Additionally, IRT true score equating results in a higher bias and RMSE than IRT observed score equating. These findings suggest that item preknowledge has the potential to inflate equated scores, putting the validity of the test scores at risk.

Keywords: cheating, item preknowledge, equating, RMSE, bias

Introduction

Testing programs often reuse items due to the difficulty and expense of creating new items. When items are reused, the security of those items poses a potential problem. Major testing organizations such as the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) define test security as "protection of content of a test from unauthorized release or use, to protect the integrity of the test scores so they are valid for their intended use" (AERA/APA/NCME, 2014, p. 236). When the security of test items is violated through preknowledge by some or all individuals taking the test, the results do not provide a valid indication of the knowledge or ability of test-takers.

Preknowledge is a form of cheating (Cizek & Wollack, 2017; Lee, 2018). Test cheating reduces the reliability and validity of the test scores (Man et al., 2019). Researchers described cheating as an ethical error (Fly, 1995) and defined it as any actions that breaches the rules of tests (Cizek, 1999). Cheating among students has increased in part due to the increased speed and straightforwardness of communication. For example, as many as 95% of university students and 53% to 60% of high school students admit to having cheated on at least one test during their educational career (Josephson Institute, 2012; Wang et al., 2015). Reports showed that some teacher candidates used systematic cheating on teacher selection exams in Turkey in 2010 and 2011 (Demir & Arcagok, 2013). What is clear from examples such as these is that cheating is a problem at all educational levels. Because of the potentially serious impact of cheating on test results, its detection is critical.

Item preknowledge occurs when examinees obtain access to test items or their answers before taking the test (Foster, 2013; Gorney & Wollack, 2022). Cizek and Wollack (2017) argue, if any cheating occurs, the resulting test scores might not accurately reflect the actual knowledge of the individuals who cheat. This means that the test may need to measure the skill or ability being assessed accurately. In

* Assoc. Prof. Dr., Ordu University, Faculty of Education, Ordu-Türkiye, akincgdm@gmail.com, ORCID ID: 0000-0001-5255-8792
** Prof. Dr., University of Georgia, College of Education, Athens-USA, acohen@uga.edu, ORCID ID: 0000-0002-8776-9378

To cite this article:
Akın-Arıkan, Ç. & Cohen, A. (2023). The impact of item preknowledge on scaling and equating: Item response theory true and observed score equating methods. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 455-471. https://doi.org/10.21031/epod.1199296

Received: 4.11.2022
Accepted: 25.10.2023

such a case, the comparability of scores across individuals and the validity of the test would be threatened (AERA et al., 2014; Lee, 2018), as exceptionally high scores might be due to prior knowledge rather than to the ability and preparedness of the examinee (Qian et al., 2016). In addition, preknowledge affects the interpretation of the test results of all individuals, not just those who had the unfair advantage of preknowledge. Further, it is the responsibility of test administrators to establish that no test-taker has an unfair advantage over others.

Using different test forms can help mitigate the effects of preknowledge, but it is then necessary to confirm that the different forms are of equal difficulty; this ensures that examinees are indifferent to which form they receive (Lord, 1980). The psychometric approach to ensuring different forms of a test are of comparable difficulty is to place these forms on the same score scale (Kolen & Brennan, 2014; von Davier et al., 2004). In that way, the scores of each form have the same interpretation.

Methods for placing test forms on the same scale are referred to as equating. If the test forms are to be equated, the first step is to decide on the equating design. In IRT-based equating, there are different designs, including random group, single group, and the non-equivalent groups with anchor item (NEAT) design. In this study, we investigate the effects of item preknowledge on equating results using the NEAT design. This design is flexible and can be used to equate multiple different test forms to a common scale. For simplicity, however, we present the following discussion in terms of equating two different forms of a test.

In the NEAT design, two groups of examinees each take one unique test form and one anchor test, in which the anchor test is the same for both groups. This anchor test is used to link the test forms to each other (Kolen & Brennan, 2014). One concern with using two different forms of a test is that the two groups of examinees may sometimes sit the exam at different times. In such a case, it is possible that information about some or all of the items on the test from the first testing group may be passed on to individuals in the other group. This creates a potential preknowledge situation for the examinees from the second group who receive the information. It is important to note that items in the anchor test are not typically identifiable as being on the anchor test. That is, examinees in the first group may not know which items are on the anchor test and which items are unique to the first test form, such that complete information may not necessarily be available to examinees from the other group. If some of the items passed along by people in the first group are anchor items, however, it could result in some members of the second group having inflated test scores. The scores of examinees with preknowledge would likely be inflated and would not accurately reflect the true abilities of these test-takers. Tan (2001) states that the results obtained from tests with individuals with preknowledge will not be valid for score-based decisions. Thus, item preknowledge can pose a serious problem in test security for test developers, test administrators, and users of test results (Pan & Wollack, 2021).

Item Response Theory

Item response theory (IRT) models are commonly utilized for test analysis and scoring (González & Wiberg, 2017). IRT models can be used for dichotomous and polytomous items. In this study, we focused on equating dichotomously scored items. IRT is used to obtain estimates of item and ability parameters. The Rasch model, one-parameter logistic (1PL) model, two-parameter logistic (2 PL) model, and three-parameter logistic (3 PL) model are among the IRT models frequently employed for analyzing dichotomously scored data. The 2 PL model, which is the model used in this study, is a generalized form of the 1PL model: the 1PL model has only item difficulty parameters, while the 2 PL model also includes item discrimination parameters. The 2 PL model takes the following mathematical form:

$$P(X_i = 1|\theta) = \frac{\exp[a_i(\theta - b_i)]}{1+\exp[a_i(\theta - b_i)]} \quad (1)$$

where θ is the individual's level of ability, a_i is the item discrimination parameter for item i, and b_i is the item difficulty parameter for item i (de Ayala, 2009).

Item Response Theory (IRT) Based Equating

In large-scale test applications, including PISA and TIMSS etc., different test forms of similar content and difficulty are generally used. Using different test forms on various dates raises concerns about potential differences in difficulty. To address this, equating is employed to make scores interchangeable between test forms, ensuring their interchangeability (Kolen & Brennan, 2014). Equating methods can be classified as traditional, kernel, local, and IRT based (which is used in the present study) equating methods (González & Wiberg, 2017).

IRT-based equating methods are classified into the following two categories: true score equating and observed score equating. In IRT, equating takes place in three stages. These steps are, respectively, estimation of item parameters, calibration, and equating the test scores. Examinees who take different test forms are not considered equivalent, and the parameters of the test forms are not represented on the same IRT scale (Kolen & Brennan, 2014). Therefore, once the item parameters have been estimated with the appropriate IRT model, the item and ability parameters can be estimated using separate or concurrent calibration, which is the first stage in the test equating. The calibration of IRT scales aims to link the new and old forms together. Through the concurrent calibration, the parameters of test forms can be estimated together, and common items are assumed to have the same parameter values in both test forms. Separate estimating methods based on test characteristics curves were shown to be the most reliable in practice (Kolen & Brennan, 2014). As a result, separate calibration methods were used in the present study.

In NEAT, items are in common from one test to the other, which allows for test forms to be linked to a common scale. However, parameter estimates from different test forms may not be on the same scale, for which a linear transformation should be performed (González & Wiberg, 2017; Kolen & Brennan, 2014). The one test is chosen as the base scale, and then the common items are used to place item parameter estimates, examinee ability estimates, and estimated ability distributions on the base scale using separate calibration methods: mean/mean (MM), mean/sigma (MS), and characteristic curve methods that are Haebara (HB) and Stocking Lord (SL). The characteristic curve methods give more consistent results for dichotomous IRT models than the mean/sigma and the mean/mean methods (Kolen & Brennan, 2014). The MS and SL methods were used in this study. The MS method is preferred because it might be easily influenced by variations in item strength, whereas the SL method is preferred because it gives more consistent results. The separate calibration can be done in the NEAT design using orthogonal regression (e.g., Kane & Mroch, 2020). For this purpose, the linking coefficients of the regression, *A* (slope) and *B* (intercept), are used. The parameters of the anchor items are used for transforming the θ-scale of form X (new form-target) to the θ-scale of form Y (old form-base form). Typically, raw scores on the new form are equated to raw scores on the old form (Kolen & Brennan, 2014). Following calibration of the items, the resulting item and ability parameter estimates are used in the equating.

The IRT true score equating method is used to link the number of correct scores on the two forms. It is done by assuming that a given θ-related true score obtained with the base scale form is equivalent to the true score of the θ in the new form. IRT observed score equating, on the other hand, uses the observed-score distributions of the two test forms obtained using the given IRT model (Han et al., 1997). These are weighted for the two distributions using equipercentile equating in IRT observed score equating (Kolen & Brennan, 2014).

There are several key differences between these two equating methods. First, IRT observed score equating specifies the equating relationship for the observed scores, while IRT true score equating uses the true scores for equating (although these are not available in practice) (Kolen & Brennan, 2014). Additionally, IRT observed score equating is sample dependent, while IRT true score equating is sample invariant (Cook & Eignor, 1991; Han et al., 1997). However, IRT true score and IRT observed score equating methods are comparable in terms of errors.

Purpose of Study

The purpose of tests is to make valid decisions about individuals in accordance with a specific aim. In order to do this, tests are expected to reflect the true ability of the individuals accurately. In other words, it is expected that highly talented individuals will score well while less talented individuals will receive lower scores. However, if individuals who take the test also have preknowledge about one or more items, they will likely score higher on those items, reducing the validity of the test (Eckerly, 2017). In IRT, the probability of answering exposed items correctly decreases as the item difficulty increases (Zimmermann et al., 2016). It can reasonably be expected that the item discrimination parameters will also change. Furthermore, the ability estimates for individuals also change with this change in item parameters, and an increasingly negative effect on the performance of honest individuals will be observed relative to the performance of individuals who cheat.

Equating can be employed to correct for differences in test form difficulty. As noted above, preknowledge among some test takers can result in corruption of the equating of form difficulties. However, it is a concern that usual equating methods do not consider item preknowledge. Thus, standard equating methods used to correct group ability differences may exacerbate the inaccuracy of the equating. It is likely that the scores obtained from the equating of tests with preknowledge among some test takers will not accurately correct for form difficulties. Therefore, it is important to determine how the presence of exposed items affects the equating.

IRT equating is a useful methodology and has been used in test construction by several testing programs and companies (Skaggs & Lissitz, 1986). However, few studies have been presented on the effects of exposed items on test equating (Barri, 2013; Jurich et al., 2010; Jurich, 2011). It would be useful, therefore, to investigate the extent to which errors in test equating change as a result of preknowledge. Barri (2013) analyzed the impact of exposed anchor items on the equated scores obtained under Rasch IRT true score equating. As the number of exposed items increased, Barri found that test scores exhibited inflated results. Similarly, Jurich (2011) investigated the effects of cheating on equated scores using 3PL true score equating for five equating methods, including the SL approach, the MM, MS method, the HB method, and the fixed anchor method. Results indicated that cheating artificially equated scores for all five methods. More recently, Liu and Becker (2022) studied the impact of item exposure and preknowledge on 1PL model pre-equated item difficulty and ability estimates. Results showed that item exposure had a significant impact on item difficulty for exposed and nonexposed items. In the previous studies (e.g., Barri, 2013; Liu & Becker, 2022) examining the effect of item preknowledge on test equating, it was seen that the 2 PL model, which is also used in large-scale test applications, hasn't been used. In addition, IRT observed score equating methods hasn't been used. In our study, we build on the previous research knowledge base by investigating the effects of preknowledge on IRT true score and IRT observed score equating methods with MS and SL. More specifically, this study aims to examine the impact of the exposure of anchor items with the 2 PL model on IRT true score and IRT observed equating under the NEAT design.

Method

A Monte Carlo simulation study was conducted to investigate the impact of exposure of anchor items. Results were compared for IRT true score equating and IRT observed score equating.

Simulation Conditions

We investigated the effects of three conditions on equating errors: ability distribution, exposed anchor items, and the proportion of examinees with pre-knowledge. The sample size is 2000, the test length is 40, and both variables were handled as constants. The simulated data was equated using the NEAT design. NEAT, the most widely used equating design, is used in Turkey in the Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and the Monitoring and Evaluation of Academic Skills (ABIDE) test administrations. Simulation conditions were listed in Table 1.

Table 1
Factors in the Simulation Design

Factor	Condition
Ability distribution (old & new forms)	(0, 1) & (.05, 1) (0, 1) & (-.2, 1.25) (0, 1) & (-1, 1)
The number of exposed anchor items	2 (20%), 6 (50%), & 10 (100%)
Proportion of examinees with preknowledge	5%, 10%, 30%, & 60%

Ability Distribution

Another important factor in equating is the ability distribution. Wang et al. (2008) suggested that a mean difference in ability of 0.05 to 0.10 is considered "relatively large," while a difference of 0.25 is considered "very large." The ability distribution of examinees who took the old form was generated using a standard normal distribution $\theta \sim N(0,1)$; however, the ability distribution of examinees who took the new form varied between conditions. In this study, three different ability distributions were analyzed: $\theta \sim N(0.05, 1.00)$ was chosen as relatively large, $\theta \sim N(-0.20, 1.25)$ was chosen as large, and $\theta \sim N(-1.00, 1.00)$ was chosen as an unacceptably large ability mean difference. A low-ability examinee had an estimated ability level of less than 0, whereas a high-ability examinee had an estimated ability level of greater than 0 (Zopluoglu, 2017). Additionally, it has been suggested that individuals of lower ability are more likely to cheat (Cizek & Wollack, 2017). Therefore, all but one of the groups simulated in this study had ability means of less than 0.

Exposed Anchor Items

Several studies have examined the effects of differing levels of preknowledge. For example, Jurich (2011) used conditions in which 5 of 10 and 10 of 10 anchor items were exposed. Barri (2013) had 2 of 10 anchor items exposed. However, it is important to note that some studies, such as Eckerly (2017), suggest that a high degree of item compromise is not typical of tests. In fact, should this occur, the validity of the scores would be seriously compromised. Bearing this in mind, the number of exposed anchor items in this study was set at 2 (20%), 6 (60%), and 10 (100%) out of 10 anchor items. In addition, a condition in which no items were exposed was included. In this way, the change that occurs as the number of items exposed increases is more clearly observed.

Proportion of Examinees with Preknowledge

A range of percentages of examinees with preknowledge have been reported in other studies. The percentages of examinees with preknowledge were determined by Barri (2013) as 5%, 10%, 15%, and 20%; Zopluoglu (2017) as 20%, 40%, and 60% and Lee (2018) as 10%, 20%, 50%, and 70%. Considering the proportions of participants who had preknowledge in other investigations, the following values were employed in this study: 0%, 5%, 10%, 30%, and 60%.

Examinees' probability of correctly responding to an exposed anchor item must also be taken into account. Previous studies have used values of .50 (Jurich, 2011), 1.00 (Barri, 2013), and/or .90 (Belov, 2016; Lee, 2018; Sinharay, 2017). In this study, the probability of a correct response was set at .90. In non-exposed conditions, no coefficient was added for the probability of correct answers to the anchor items. In other terms, the response probability in the non- exposed condition was modeled by the 2 PL model. The non-exposed condition was used as a basis for the comparisons. This situation was created as a situation where anchor test items were not shared between the groups in the test application with the NEAT.

Data Generation

The dichotomous item response data under the 2 PL model were generated using R (R Core Team, 2021). In the previous research, the Rasch model (Barri, 2013), the 3PL model (Jurich, 2011), and the 1PL model (Liu and Becker, 2022) were used to examine exposed items on test equating. For this reason, the 2 PL model, which is also preferred in large-scale test applications such as PISA, was used in this

study. The latent trait model (ltm) package (Rizopoulos, 2006) was used for item and ability parameter estimations. This package provides marginal maximum likelihood for item parameters estimation and expected *a posteriori* for ability parameters estimation. The NEAT design requires two different test forms (old and new) with an anchor test, and also two different groups taking one test.

Test length: Spence (1996) has suggested that tests should have a minimum length of 35 items for equating purposes, whereas Kolen and Brennan (2014) suggested a range of 30 to 40 items. For this particular study, both the old and new versions of the test were constructed with a total of 40 items. Angoff (1984) and Kolen and Brennan (2014) suggest that the number of anchor items should be 20% of the test. Based on these suggestions, the anchor test was set at 10 items. This study utilized a test length of 40, with 10 internal anchor items used for both forms; meeting all of the above criteria for test length.

Sample Size: Kolen and Brennan (2014) noted that random equating error is influenced by sample size and suggest a minimum sample size of 400 per test form for linear equating methods and of 1,500 for equipercentile equating (Harris, 1993; Kolen & Brennan, 2014). Spence (1996) similarly recommends a minimum sample size of 500 for accurate equating results. In this research, a sample size of 2,000 was used, which exceeds each of these recommendations.

Ability distributions: For each group, the ability distribution was sampled from a standard normal distribution for all conditions.

Item parameters: Item difficulties were generated using a random normal distribution with the rnorm function, and item discrimination parameters were generated using a random log-normal distribution with the rlnorm function. For the old form and the anchor test, item difficulties had a mean of 0.00 and a variance of 1.00; the new form had a mean of 0.05 and a variance of 1.00. It was thought that the difference of .05 represented two forms with similar difficulties. The item discrimination parameter had a mean of 0.30 and a variance of 0.20. Descriptive statistics for the generated parameters are given in Table 2.

Table 2

Descriptive Statistics of True Item Parameters

Descriptive statistics	*b* (Old Form)	*b* (New Form)	*a* (Old Form)	*a* (New Form)	*b* (Anchor test)	*a* (Anchor test)
Mean	0.08	0.12	1.48	1.38	-0.08	1.47
SD	0.93	1.09	0.36	0.30	1.44	0.39
Min.	-1.81	-1.78	0.93	1.03	-2.02	1.06
Max	1.91	2.42	2.09	2.08	2.12	2.09

SD: standard deviation

As shown in Table 2, the *b* parameters (i.e., item difficulty parameters) for the old form ranged from -1.81 to 1.91, with a mean of 0.08 and an SD of 0.93. The *b* parameters for the new form ranged from -1.78 to 2.42, with a mean of 0.12 and an SD of 1.09. This design includes a noticeable difference in the mean difficulties of the old and new forms, but the *a* parameter values for the two were similar. For anchor items, *b* parameters ranged from -2.02 to 2.12, and *a* parameters ranged from 1.06 to 2.09.

Scaling Methods

The *plink* package (Weeks, 2010) in R was used for scaling transformations with MS and SL and equating tests under IRT true and observed score test equating.

The MS method (Marco, 1977) is the scaling method that uses the means and standard deviations of the *b* parameters of the common items to estimate the linking coefficients: *slope (A)* and *intercept (B)* in IRT scale transformation. The mean of item parameters $\mu(b_J)$, $\mu(b_I)$, $\sigma(b_J)$, and $\sigma(b_I)$ are given below:

$$A = (\sigma(b_J)) \,/(\sigma(b_I)) \qquad (2)$$

$$B = \mu(b_J) \; - \; A\mu(b_I) \qquad (3)$$

The other method used in this study was the SL characteristic curve method (Stocking & Lord, 1983), which is one of the most widely used IRT-based equating. This is done by applying summation to the parameter estimates before squaring. The SL equation can be given as follows:

$$SL_{diff}(\theta_i) = \left[\sum_{j:V} \theta_{Ji}; \hat{a}_{Ji}, \hat{b}_{Ji} + \hat{c}_{Ji} \sum_{j:V} p_{ij} \left(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\,\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)\right]^2 \quad (4)$$

The A and B coefficients obtained by using the MS and SL methods can then be used to transform the θ and item parameter estimates to the base scale as follows:

$$\theta_{Ji} = A\theta_{li} + B \quad (5)$$

$$a_{Ji} = \frac{a_{ti}}{B} \quad (6)$$

$$b_{Ji} = Ab_{li} + B \quad (7)$$

After calibration, the equating step was performed using IRT true score (IRT-T) and observed score (IRT-O) equating methods.

In IRT true score equating, the old test, $X(\theta)$, and the new test, $Y(\theta)$, are regarded as equivalent for a given θ. The true score of θ_i*:* is indicated by τx^{-1}

$$\tau(X) = \tau(Y)(\tau x^{-1}) \quad (8)$$

IRT true score equating has the following three steps, and each step is performed for all true scores (Kolen & Brennan, 2014):

1. Choose a true score from form X [$\tau(X)$].
2. Identify the θ_i *corresponding* to the true score.
3. Define the true score of form Y that corresponds to θ_i.

In IRT observed score equating, after the observed score distribution of each form is obtained using IRT models, the tests are equated using the equipercentile method. The IRT observed score equating method consists of four steps (Kolen & Brennan, 2014):

1. For forms X and Y, the distribution of observed scores is calculated using the compound binomial distribution for examinees of a given ability. This is done using the recursion formula.
2. The distribution of observed examinee scores at each ability is obtained using equations 9 through 12 for forms X and Y. The distributions are then added together.

$$f_1(x) = \sum_i f(x|\theta_i)\varphi_1(\theta_i) \quad (9)$$

$$f_2(x) = \sum_i f(x|\theta_i)\varphi_2(\theta_i) \quad (10)$$

$$g_1(x) = \sum_i g(y|\theta_i)\varphi_1(\theta_i) \quad (11)$$

$$g_2(x) = \sum_i g(y|\theta_i)\varphi_2(\theta_i) \quad (12)$$

3. For IRT observed score equating under the NEAT design involving two populations, an equating function is typically viewed as defining a single population. Thus, populations 1 and 2 must be equated to be able to treat them as a single population. Populations 1 and 2 are weighted by w_1 and w_2 to form a synthetic population where $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$. Synthetic weights are used to determine the distributions in the synthetic population.

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x) \quad (13)$$

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y) \quad (14)$$

4. The traditional equipercentile method is used to obtain equated scores. For this study, synthetic population weights of .50 were used for both populations for all groups ($w_1 = w_2 = .50$) for IRT observed score equating. The use of equal weights means that both populations were treated as contributing equally to the synthetic population (Kolen & Brennan, 2014).

Recovery in IRT Equating

The aim of this research is to see how the ability distribution, number of exposed anchor items, and proportion of examinees having preknowledge affect test equating under the 2 PL model. Root mean square error (RMSE) and bias were calculated to evaluate the recovery of the equating scores and scaling coefficients for the slope and intercept.

The equations for calculating bias and RMSE are given below:

$$Bias(x) = \frac{1}{R}\sum_{j=1}^{n} \hat{e}_y(x) - e_y(x) \quad (15)$$

$$\text{RMSE (x)} = \sqrt{\frac{\sum_{j=1}^{n}((\hat{e}_y(x) - e_y(x))^2}{R}} \quad (16)$$

where R is the number of replications, $e_y(x)$ is the true value, and $\hat{e}_y(x)$is the estimated value of each replication. 100 replications of the data were generated for each combination of ability distributions, the number of exposed anchor items, and the proportion of examinees with preknowledge.

Results

This simulation study was conducted to evaluate the impact of exposed anchor items on IRT true score observed equating methods under the 2 PL model.

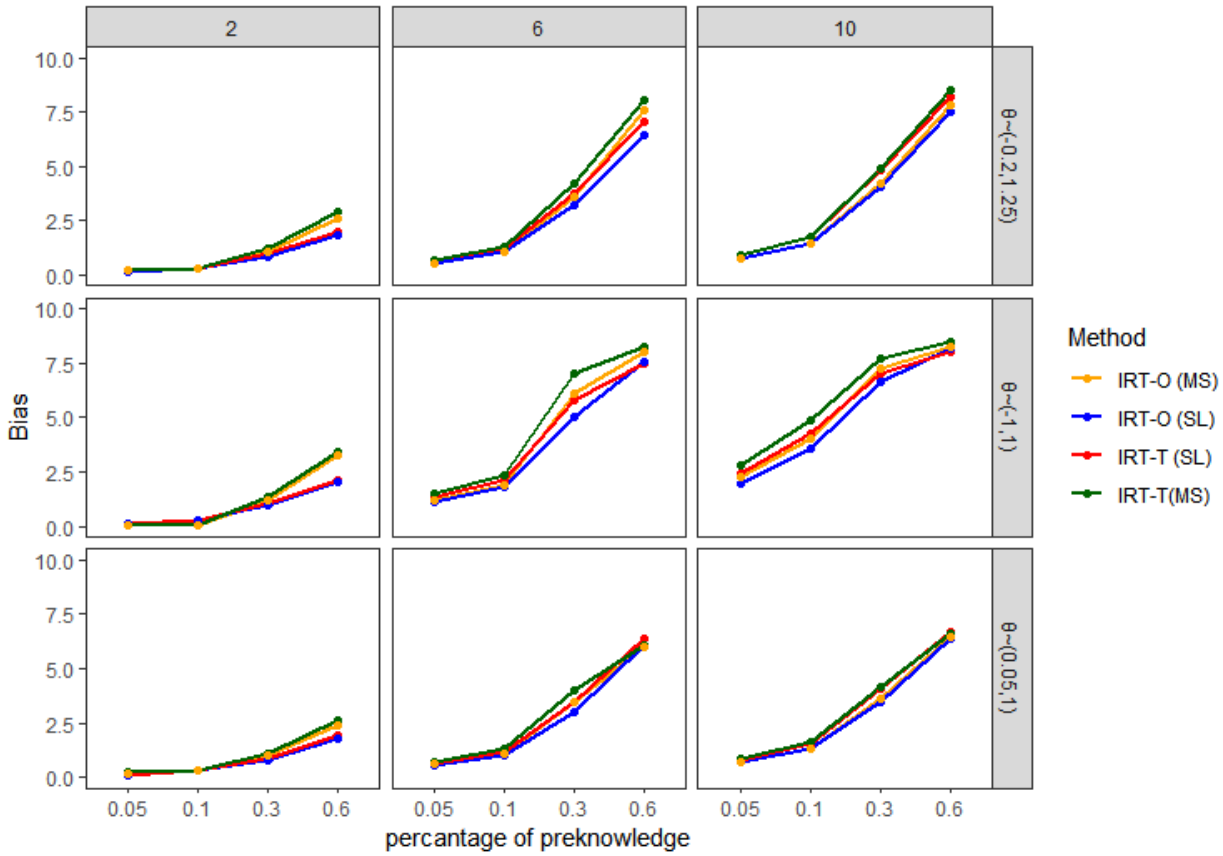
Bias and RMSE of Equated Scores

Figure 1 shows bias results for equating under different numbers of exposed anchor items and percentages of examinees with preknowledge for each of the different mean ability distributions. The bias under the nonexposed condition was close to zero (see Appendix 1) but positive in all exposure conditions. Positive bias indicates that the examinees with preknowledge produced higher scores than expected for the given condition. Bias increased slightly for each scaling method for the two exposed anchor items in both the true score and observed score equating methods compared to the condition with nonexposed item. The condition with two exposed items had a similar increasing pattern for ability distributions except for θ~N(-1,1). The MS scaling method showed a higher bias than the SL method when preknowledge was set at 60%. For the condition with two exposed anchor items, the largest amount of bias was found for the 30% condition, though bias was also observed for the 10% condition when the number of exposed items was six and 10.

The condition with 10 exposed items and 60% preknowledge resulted in larger, positively biased equated scores under IRT true score equating with the MS scaling method. As the number of exposed anchor items increased, bias also increased for both scaling methods under both equating methods. SL performed the best and produced the least bias for IRT observed score equating methods under the θ~N(0.05,1) ability condition. Both equating methods had similar amounts of bias under the θ~N(0.05,1) ability condition for all numbers of exposed items. For true score equating with the MS scaling method, the ability distribution θ~N(-1,1) produced the largest amount of bias except for the condition with two exposed anchor items. For observed score equating with SL scaling, the ability distributions θ~N(0.05,1) and θ~N(-0.2,1.25) produced the least bias.

Figure 1

Bias of Equated Score under IRT True and Observed Score Equating Methods with Different Scaling Methods



Overall, IRT true score equating produced higher levels of bias for all conditions than did IRT observed score equating. Additionally, the MS method produced more biased scores than the SL method. The largest amount of bias was observed when using the MS method under IRT true score equating, with individual scores being an average of 8.46 raw score points above the expected scores. It is also clear from our results that the ability distribution affected the estimated equating scores and that the number of exposed items affected the accuracy of equating.

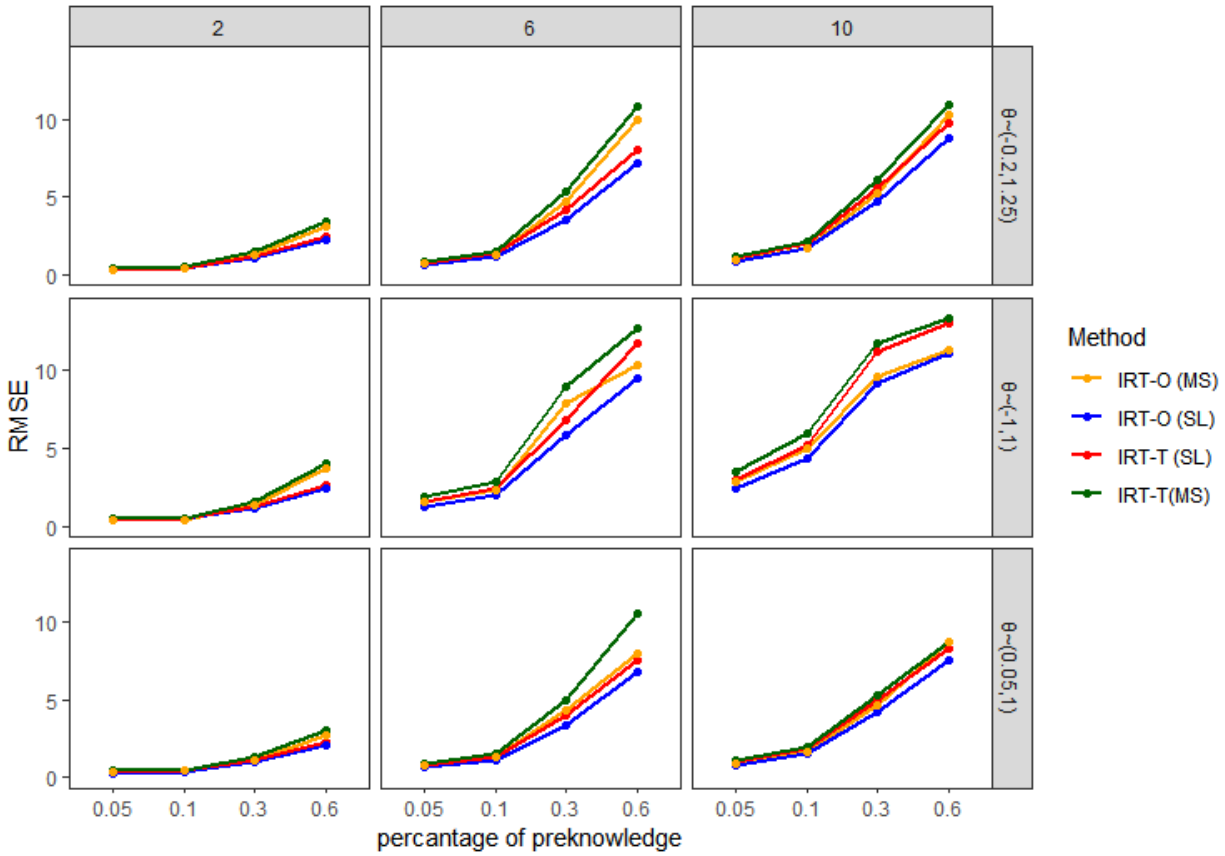
Figure 2 shows the RMSE results for equating scores under different conditions. The RMSE results for both equating methods had the smallest values under the nonexposed conditions, whereas the largest value was obtained from the MS scaling method under IRT true score equating with the ability distribution $\theta \sim N(-1,1)$ (see Appendix 2).

In the condition with two exposed anchor items and 5% to 30% of examinees with preknowledge, the RMSE increased slightly for each scaling method. The RMSE for the MS scaling method was higher than that of the SL scaling method when the ability distribution was $\theta \sim N(-1,1)$, and the level of preknowledge was less than 60%. When the percentage of examinees with preknowledge increased from 5 to 10%, the RMSE increased. However, when preknowledge was increased from 10 to 30% and from 30 to 60%, the RMSE approximately doubled. This is especially noticeable for the MS scaling method with six exposed anchor items; the highest increase was from 30 to 60%. In addition, when the ability distribution was $\theta \sim N(-1.1)$, the IRT true score equating method with SL and MS exhibited a large RMSE with 60% preknowledge. When the ability distribution was $\theta \sim N(-0.2.1.25)$, the MS scaling method had a high RMSE for both equating methods. When there were 10 exposed anchor items with 5 to 10% preknowledge, the RMSE increased slightly for each scaling method when the ability distributions were $\theta \sim N(0.05,1)$ and $\theta \sim N(-0.2.1.25)$; however, the RMSE was particularly high, when the preknowledge of

examinees was between 10 and 60% for θ~*N*(-0.2.1.25). In addition, when the mean ability distribution of the groups was different [θ~*N*(-0.2.1.25) vs. θ~*N*(-1.1)], the discrepancy between IRT true and observed score equating methods increased. This differentiation was most obvious when preknowledge was set at 60%. Another finding is that when the mean of the ability distribution was negative, the number of exposed items was six, and preknowledge was set at 60% (or when the number of exposed items was 10), IRT true score equating with the SL and MS scaling methods gave a higher RMSE than did IRT observed score equating.

Figure 2

RMSE of Equated Score under IRT True and Observed Score Equating Methods with Different Scaling Methods



As the number of exposed anchor items increased, the RMSE also increased under all conditions. The largest RMSE value was obtained from the ability distribution θ~*N*(-1.1) using the MS scaling method under IRT true score equating; in contrast, the smallest RMSE was produced from the ability distribution θ~*N*(0.05,1) with the SL scaling method under IRT observed score equating.

Bias and RMSE of Slope and Intercepts of the Scaling Methods

Table 3 shows that bias under nonexposed conditions was nearly zero for both the slope and the intercept. The bias of the slope was low when the percentage of preknowledge was 5 or 10%; however, this bias increased when the percentage of preknowledge reached 30% and then 60%. On the other hand, the bias of the intercept was still close to that of the nonexposed conditions when the percentage of preknowledge was 5 or 10% with two exposed items. As the percentage of preknowledge and the number of exposed items increased, the bias also increased. Additionally, both scaling methods underestimated the slope when two items were exposed. In contrast, as the number of exposed items increased, the slope was overestimated for all ability distributions when six or 10 items were exposed. The intercept, on the other hand, was overestimated by both scaling methods for nearly all conditions,

and the recovery of the slope was more accurate than that of the intercept in almost all conditions. The results show that the recovery of the slope and intercept was affected by the ability distribution: the most biased estimate of the slope was obtained using ability distribution of θ~*N*(-1,1), 60% preknowledge, and 10 exposed items under the MS scaling method. The MS method had a larger bias than the SL for all conditions, except with 10 exposed items and 60% preknowledge when the mean ability distribution was θ~*N*(-1,1).

Table 3

Bias of Slope and Intercept

		Ability Distribution											
		θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
Scaling methods		SL		MS		SL		MS		SL		MS	
Item Pre.	Perc. of Pre.	A	B	A	B	A	B	A	B	A	B	A	B
none		0.01	-0.01	0.01	-0.01	-0.01	0.01	-0.01	0.00	0.01	0.02	0.02	0.03
2	5%	0.02	-0.03	0.01	-0.04	0.00	-0.02	-0.01	-0.02	0.03	0.02	0.02	0.04
	10%	0.03	-0.07	0.01	-0.07	0.01	-0.05	-0.01	-0.04	0.05	-0.01	0.04	0.03
	30%	0.05	-0.20	0.02	-0.20	0.04	-0.16	-0.03	-0.16	0.09	-0.18	0.05	-0.18
	60%	0.11	-0.45	0.11	-0.55	0.14	-0.38	-0.12	-0.48	0.22	-0.46	0.22	-0.69
6	5%	-0.02	-0.12	-0.04	-0.12	-0.01	-0.09	-0.04	-0.08	-0.05	-0.17	-0.11	-0.17
	10%	-0.03	-0.21	-0.07	-0.20	-0.01	-0.17	-0.07	-0.15	-0.05	-0.31	-0.17	-0.29
	30%	-0.10	-0.62	-0.28	-0.60	-0.03	-0.55	-0.25	-0.51	-0.24	-0.92	-0.51	-0.98
	60%	-0.11	-1.14	-0.59	-1.33	0.03	-1.20	-0.50	-1.26	-0.48	-1.63	-0.54	-1.72
10	5%	-0.04	-0.13	-0.05	-0.14	-0.03	-0.11	-0.05	-0.11	-0.16	-0.30	-0.21	-0.33
	10%	-0.08	-0.25	-0.11	-0.25	-0.05	-0.23	-0.09	-0.21	-0.28	-0.55	-0.34	-0.60
	30%	-0.25	-0.66	-0.32	-0.65	-0.15	-0.65	-0.25	-0.62	-0.65	-1.11	-0.62	-1.23
	60%	-0.50	-1.15	-0.56	-1.16	-0.27	-1.30	-0.45	-1.27	-0.81	-1.80	-0.78	-1.87

A: slope; B: intercept; Perc.of Pre: percentage of preknowledge

Table 4

RMSE for Slope and Intercept

		Ability Distribution											
		θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
Scaling methods		SL		MS		SL		MS		SL		MS	
Item Pre.	Perc. of Pre.	A	B	A	B	A	B	A	B	A	B	A	B
none		0.02	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.05
2	5%	0.04	0.04	0.03	0.05	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.05
	10%	0.04	0.08	0.03	0.08	0.02	0.05	0.03	0.05	0.06	0.05	0.06	0.06
	30%	0.06	0.20	0.04	0.20	0.05	0.16	0.04	0.17	0.10	0.18	0.08	0.19
	60%	0.11	0.45	0.12	0.55	0.14	0.38	0.13	0.49	0.22	0.47	0.22	0.70
6	5%	0.03	0.12	0.05	0.13	0.02	0.09	0.05	0.08	0.05	0.17	0.12	0.17
	10%	0.04	0.21	0.09	0.20	0.02	0.18	0.08	0.15	0.06	0.31	0.18	0.29
	30%	0.11	0.62	0.29	0.61	0.04	0.56	0.26	0.51	0.25	0.93	0.51	0.99
	60%	0.12	1.14	0.59	1.14	0.04	1.25	0.50	1.26	0.50	1.64	0.54	1.74
10	5%	0.05	0.13	0.07	0.14	0.04	0.11	0.06	0.11	0.16	0.30	0.21	0.34
	10%	0.08	0.25	0.13	0.26	0.05	0.23	0.09	0.23	0.29	0.55	0.35	0.61
	30%	0.25	0.66	0.33	0.66	0.15	0.65	0.26	0.63	0.65	1.12	0.63	1.24
	60%	0.50	1.16	0.59	1.17	0.28	1.31	0.49	1.28	0.81	1.81	0.79	1.88

A: slope; B: intercept; Perc.of Pre: percentage of preknowledge

Table 4 shows the RMSE for the slope and intercept under various conditions. The RMSE under nonexposed conditions was nearly zero for both the slope and the intercept. At 5 and 10% preknowledge, the RMSE was close to that of the nonexposed conditions; however, the RMSE increased for 30% preknowledge, and increased again for 60%. These results suggest that the recovery of the slope was affected by the ability distribution. The largest RMSE of the slope was obtained when the ability distribution was $\theta \sim N(-1,1)$, preknowledge was 30% or 60%, and 10 items were exposed under the SL scaling method. As with the slope, the RMSE of the intercept increased when the percentage of preknowledge and number of exposed items increased. Under all conditions, however, the recovery of the slope was more accurate than that of the intercept. The SL and MS scaling methods had a similar RMSE for all conditions, except when preknowledge was set at 60%. The RMSE was close to that of nonexposed conditions for 5 and 10% preknowledge with two exposed items, though it was larger for 30% and 60% preknowledge. In addition, these results suggest that the conditions chosen influenced the estimation of the intercept, especially the number of exposed items and mean ability distribution.

Conclusion and Discussion

In this study, we aimed to investigate the effect of the anchor item preknowledge on equated scores and scaling coefficients under IRT true and IRT observed score equating. This was premised on the idea that the validity of inferences based on test scores becomes questionable if individuals have preknowledge of the anchor items on tests.

The results of this study suggest that as the number of exposed items and percentage of examinees with item preknowledge increase, bias also increases. As all bias observed in this study was positive, the equated scores were estimated with higher values than the true (i.e., generating) values. The amount of bias differed based on the scaling and equating methods used. Results obtained from MS exhibited a larger bias than those obtained from SL. Our finding that the SL method provides more accurate and stable equating results than the MS method is in line with previous research (Kim & Cohen, 1992; Kim & Kolen, 2006; Kim & Lee, 2006). It can be explained that MS, which requires simple summary statistics (Kim, 2004), has higher errors because it is more sensitive to the variation of the estimates of the b parameter, and as a result, the slope and the intercept values may become unstable. Furthermore, bias and RMSE increased with the number of exposed items and percentage of preknowledge, which is also consistent with previous research (Barri, 2012; Chen, 2021; Jurich, 2011; Kopp & Jones, 2020).

Both linking methods had higher bias values for IRT equating than the nonexposed condition. However, results for 5 and 10% preknowledge for the two exposed items condition had bias values close to those of the nonexposed condition. One possible reason for this may be that the MS method considers item parameters separately while the SL method considers them simultaneously (Kolen & Brennan, 2014; Tian, 2011). The MS method is more directly affected by variation in the item difficulty parameter since the scaling coefficients depend on the item difficulty parameter, and item preknowledge increases the probability of a correct answer. This result is consistent with the findings of Lee and Becker (2022), who reported that as the percentage of examinees with preknowledge increases, the variance of the item difficulty parameter estimates increases for exposed conditions. Finally, consistent with previous research (Barri, 2012; Jurich, 2011), when the ability distributions were similar or equivalent, the bias and RMSE were lower and thus appeared to have a more minor effect on equated scores.

IRT true score equating exhibits a larger bias and RMSE does IRT observed score equating for both scaling methods. However, bias and RMSE values for both equating methods were similar in the nonexposed condition. Our findings are consistent with Tao and Cao's (2016) findings, which showed that IRT observed score equating outperforms the IRT true score equating. However, others found that IRT observed score equating are more stable compared to IRT true score equating (Han et al., 1997). Due to different results on equating methods in the related literature, there is no consensus on the best method. IRT observed score equating uses synthetic weights, while IRT true score equating uses the true score to equate through an ability parameter (Ogasawara, 2003). As a result, the presence of exposed items changes the probability of a correct answer, resulting in higher scores and thereby higher bias and RMSE values. In the present study, we utilized equal synthetic weights for groups, which may have affected the difference between the equated scores of the two groups.

Higher levels of bias and RMSE were observed as item exposure and percentage of knowledge increased for both scaling methods, which is consistent with the previous research (Barri, 2012; Chen, 2021; Jurich, 2011). The scaling coefficient *A* was overestimated by both scaling methods for the condition with two exposed items and underestimated for the conditions with six and ten exposed items. In other words, the number of exposed items appeared to affect the estimation of coefficient *A*. These findings were consistent with Barri (2012) and Chen (2021) for the condition with two exposed items and with Jurich (2011) in the sense that coefficient *A* was underestimated by both estimation methods for the conditions with six and ten exposed items. The reason for these differences in the effect on scaling coefficient *A* may be that the variance of the difficulty parameter changes as the difficulty of the exposed items decreases. Jurich (2011) suggests that while the probability of a correct answer will increase as the number of exposed items increases, this may also lead to a decrease in item discrimination. This situation may also cause an underestimation of coefficient *A*. On the other hand, scaling constant *B* was underestimated under all conditions, which is consistent with Barri (2011) but not with other studies (e.g., Jurich, 2011). This disagreement may be due, in part, to the way in which item preknowledge was simulated. Barri (2012) simulated item preknowledge by adding 1 to the probability of a correct answer, but Jurich (2011) added .5. In this study, item preknowledge was simulated by adding .9. In addition, scaling coefficient A was more accurate than scaling coefficient B for both linking methods. Thus, as the number of exposed items, the percentage of examinees with item preknowledge, and the differences in mean ability increased, bias and RMSE values increased for both scaling coefficients.

Ability also had an impact on linking and equating results when item exposure occurred, which is consistent with the previous research (Barri, 2012; Chen, 2021; Jurich, 2011). As the difference in mean ability increased, bias and RMSE also increased. This effect can be seen in low-ability examinees, who correctly answer exposed items at a higher rate than higher-ability examinees (Barri, 2012).

We found that the choice of scaling or equating methods may be of less importance when items are exposed. However, the generalizability of our findings needs to be critically evaluated. Some examinees exhibit a greater change in equated scores when they have preknowledge of the anchor items. In fact, this situation affects not only examinees with preknowledge but also the decisions made about all examinees who take the test (Jurich, 2011). For this reason, the effects of exposure should be examined before equating. Otherwise, the validity of the decisions made may be open to question. The bias results from this study suggest that if the anchor items are exposed, it is most appropriate to exclude them from the test before equating, as item preknowledge affects the equated test scores.

In this study, we focused on determining the effect of item preknowledge on IRT test equating methods under the NEAT design. We acknowledge that several additional avenues for future research exist. First, in this study, equal synthetic weights were used. Future research might choose differing synthetic weights to determine their effect on equated scores. Second, examining the effect of item preknowledge on tests with mixed item formats would be useful. Third, although the NEAT design is frequently reported in the existing literature, other equating designs, such as the random group design and common-item equivalent groups design, could be used. Fourth, the variance of the item difficulty parameter estimates may differ in real data. Therefore, the effect of exposed items on the test equating can be examined in a real data set. Finally, future studies can extend our work of using IRT models to estimate parameters and equating by applying other methods (e.g., classical equating, Bayesian nonparametric, and kernel equating).

Declarations

Author Contribution: Cigdem Akin Arikan: Conceptualization, methodology, analysis, writing & editing, visualization. Allan S. Cohen: Conceptualization, writing-review & editing, supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Ethical rules were followed in this research. Ethical approval is not required, because simulation data was used in this research.

Funding: This study was funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) BIDEB 2219.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Barri, M. A. (2013). *The impact anchor item exposure on mean/sigma linking And IRT true score equating under the neat design* [Unpublished master's thesis]. University of Kansas.

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83-97. https://doi.org/10.1177/0146621615603

Chen, D. F. (2021). *Impact of item parameter drift on IRT linking methods* [Unpublished doctoral thesis]. The University of North Carolina.

Cizek, G. (1999). *Cheating on tests: how to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational measurement: Issues and practice*, *10*(3), 37-45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

de Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

Demir, M. K., & Arcagok, S. (2013). Primary school teacher candidates' opinions on cheating in exams. *Erzincan University Faculty of Education Journal*, *15*(1), 148-165. Retrieved from https://dergipark.org.tr/en/pub/erziefd/issue/6010/80121

Eckerly, C. A. (2017). Detecting preknowledge and item compromise. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 101-123). Routledge.

Fly, B. J. (1995). *A study of ethical behaviour of students in graduate training programs in psychology* [Unpublished doctoral thesis]. University of Denver.

Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack and J. J. Fremer , Eds., *Handbook of test security* (pp. 39–83). Routledge

Gorney, K., & Wollack, J. A. (2022). Generating models for item preknowledge. *Journal of Educational Measurement*, *59*(1), 22-42. https://doi.org/10.1111/jedm.12309

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*(2), 105-121. https://doi.org/10.1207/s15324818ame1002_1

Harris, D. J. (1993, April). *Practical issues in equating* [Paper presentation]. American Educational Research Association, Atlanta, Georgia, USA.

Josephson Institute (2012). *Josephson Institute's 2012 report card on the ethics of American youth*. Los Angeles, CA. Retrieved from http://charactercounts.org/programs/reportcard/2012/index.html.

Jurich, D. P. (2011). *The impact of cheating on IRT equating under the non-equivalent anchor test design* [Unpublished master's thesis]. James Madison University.

Jurich, D. P., Goodman, J. T., & Becker, K. A. (2010). *Assessment of various equating methods: Impact on the pass-fail status of cheaters and non-cheaters*. In Poster presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Kane, M. T., & Mroch, A. A. (2020). Orthogonal Regression, the Cleary Criterion, and Lord's Paradox: Asking the Right Questions. *ETS Research Report Series*, *2020*(1), 1-24. https://doi.org/10.1002/ets2.12298

Kim, S. (2004). *Unidimensional IRT scale linking procedures for mixed-format tests and their robustness to multidimensionality [*Doctoral dissertation]. Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3129309)

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed format tests. *Applied Measurement in Education*, *19*, 357–381.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of educational measurement*, *29*(1), 51-66.

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

Kopp, J. P., & Jones, A. T. (2020). Impact of item parameter drift on Rasch scale stability in small samples over multiple administrations. *Applied Measurement in Education*, *33*(1), 24-33.

Liu, J., & Becker, K. (2022). The Impact of cheating on score comparability via pool-based IRT pre-equating. *Journal of Educational Measurement*, *59*(2), 208-230. https://doi.org/10.1111/jedm.12321

Lee, S. Y. (2018). *A mixture model approach to detect examinees with item preknowledge* [Unpublished doctoral dissertation]. The University of Wisconsin-Madison.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, *56*(2), 251-279. https://doi.org/10.1111/jedm.12208

Marco, G. L. (1977). Item Characteristic Curve Solutions to Three Intractable Testing Problems. *Journal of Educational Measurement, 14* (2), 139-160.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38-47. https://doi.org/10.1111/emip.12102

Pan, Y., & Wollack, J. A. (2021). An unsupervised-learning based approach to compromised items detection. *Journal of Educational Measurement, 58*(3), 413-433. https://doi.org/10.1111/jedm.12299

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Rizopoulos, D. (2006). ltm: An R Package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46-68. https://doi.org/10.3102/1076998616673872

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529.

Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* [Unpublished doctoral dissertation]. University of Florida.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Tan, Ş. (2001). Sınavlarda kopya çekmeyi önlemeye yönelik önlemler [Measures against cheating in exams]. *Education and Science*, *26*(122), 32-40.

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education*, *29*(2), 108-121.

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

Wang, J., Tong, Y., Ling, M., Zhang, A., Hao, L., & Li, X. (2015). Analysis on test cheating and its solutions based on extenics and information technology. *Procedia Computer Science*, *55*, 1009-1014. https://doi.org/10.1016/j.procs.2015.07.1024

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*, 632-651. https://doi.org/10.1177/0146621608314943

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1–33. https://doi.org/10.18637/jss.v035.i12

Zimmermann, S., Klusmann, D., & Hampe, W. (2016). Are exam questions known in advance? Using local dependence to detect cheating. *PloS One, 11*(12). https://doi.org/10.1371/journal.pone.0167545

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance. Understanding the status Quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 25–46). Routledge.

Appendix 1

Table 1

The Bias of Equated Scores

Item Pre.	Percentage of preknowledge	Ability Distribution											
		θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
		SL		MS		SL		MS		SL		MS	
		IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O
2	non	0.02	0.03	0.05	0.06	0.02	0.03	0.05	0.05	0.00	0.02	-0.04	-0.01
	5%	0.13	0.13	0.22	0.21	0.19	0.17	0.21	0.19	0.11	0.14	0.04	0.05
	10%	0.32	0.29	0.32	0.29	0.33	0.30	0.33	0.29	0.31	0.30	0.07	0.08
	30%	0.89	0.82	1.10	0.98	0.96	0.87	1.22	1.07	1.04	0.98	1.35	1.18
	60%	1.91	1.79	2.64	2.43	2.01	1.86	2.88	2.61	2.11	2.06	3.45	3.24
6	5%	0.65	0.56	0.73	0.63	0.67	0.56	0.68	0.56	1.32	1.10	1.51	1.22
	10%	1.19	1.01	1.31	1.11	1.24	1.04	1.26	1.03	2.13	1.77	2.32	1.86
	30%	3.49	3.03	3.97	3.44	3.75	3.20	4.24	3.60	5.77	5.02	7.03	6.08
	60%	6.42	6.05	6.05	6.03	7.05	6.40	8.06	7.60	7.47	7.59	8.24	8.00
10	5%	0.79	0.68	0.87	0.74	0.90	0.75	0.92	0.76	2.41	1.98	2.82	2.29
	10%	1.52	1.30	1.60	1.35	1.74	1.46	1.72	1.42	4.28	3.54	4.84	4.00
	30%	4.05	3.50	4.18	3.61	4.79	4.09	4.89	4.17	7.02	6.65	7.70	7.28
	60%	6.68	6.39	6.61	6.43	8.23	7.49	8.46	7.84	7.99	8.14	8.46	8.27

Appendix 2

Table 2

The RMSE of Equated Scores

Item Pre.	Percentage of preknowledge	Ability Distribution											
		θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
		SL		MS		SL		MS		SL		MS	
		IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O
2	non	0.22	0.2	0.26	0.24	0.23	0.21	0.28	0.26	0.29	0.27	0.44	0.41
	5%	0.29	0.25	0.4	0.35	0.34	0.29	0.39	0.34	0.41	0.34	0.49	0.41
	10%	0.43	0.37	0.48	0.42	0.46	0.40	0.48	0.41	0.53	0.45	0.51	0.42
	30%	1.06	0.95	1.31	1.14	1.18	1.05	1.45	1.25	1.28	1.16	1.60	1.37
	60%	2.28	2.09	3.07	2.75	2.478	2.27	3.42	3.05	2.63	2.46	4.07	3.70
6	5%	0.78	0.65	0.9	0.77	0.779	0.64	0.84	0.70	1.58	1.25	1.90	1.52
	10%	1.35	1.13	1.58	1.34	1.4	1.14	1.52	1.28	2.40	1.97	2.83	2.35
	30%	3.95	3.39	4.94	4.32	4.203	3.56	5.35	4.68	6.85	5.84	9.01	7.86
	60%	7.53	6.78	10.54	7.95	8.056	7.20	10.80	10.00	11.81	9.49	12.70	10.39
10	5%	0.98	0.82	1.09	0.91	1.07	0.88	1.12	0.93	2.92	2.40	3.47	2.86
	10%	1.82	1.52	1.98	1.66	2.017	1.66	2.09	1.73	5.24	4.36	6.01	5.03
	30%	4.92	4.23	5.32	4.61	5.597	4.74	6.06	5.21	11.28	9.18	11.78	9.59
	60%	8.25	7.51	8.7	8.69	9.791	8.82	11.93	10.27	13.10	11.10	13.43	11.35