World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:12, No:4, 2018

Comparing Test Equating by Item Response Theory and Raw Score Methods with Small Sample Sizes on a Study of the ARTé: Mecenas Learning Game

Steven W. Carruthers

***Abstract*—**The purpose of the present research is to equate two test forms as part of a study to evaluate the educational effectiveness of the ARTé: Mecenas art history learning game. The researcher applied Item Response Theory (IRT) procedures to calculate item, test, and mean-sigma equating parameters. With the sample size n=134, test parameters indicated "good" model fit but low Test Information Functions and more acute than expected equating parameters. Therefore, the researcher applied equipercentile equating and linear equating to raw scores and compared the equated form parameters and effect sizes from each method. Item scaling in IRT enables the researcher to select a subset of well-discriminating items. The mean-sigma step produces a mean-slope adjustment from the anchor items, which was used to scale the score on the new form (Form R) to the reference form (Form Q) scale. In equipercentile equating, scores are adjusted to align the proportion of scores in each quintile segment. Linear equating produces a mean-slope adjustment, which was applied to all core items on the new form. The study followed a quasi-experimental design with purposeful sampling of students enrolled in a college level art history course (n=134) and counterbalancing design to distribute both forms on the pre- and post-tests. The Experimental Group (n=82) was asked to play ARTé: Mecenas online and complete Level 4 of the game within a two-week period; 37 participants completed Level 4. Over the same period, the Control Group (n=52) did not play the game. The researcher examined between group differences from post-test scores on test Form Q and Form R by full-factorial Two-Way ANOVA. The raw score analysis indicated a 1.29% direct effect of form, which was statistically non-significant but may be practically significant. The researcher repeated the between group differences analysis with all three equating methods. For the IRT mean-sigma adjusted scores, form had a direct effect of 8.39%. Mean-sigma equating with a small sample may have resulted in inaccurate equating parameters. Equipercentile equating aligned test means and standard deviations, but resultant skewness and kurtosis worsened compared to raw score parameters. Form had a 3.18% direct effect. Linear equating produced the lowest Form effect, approaching 0%. Using linearly equated scores, the researcher conducted an ANCOVA to examine the effect size in terms of prior knowledge. The between group effect size for the Control Group versus Experimental Group participants who completed the game was 14.39% with a 4.77% effect size attributed to pre-test score. Playing and completing the game increased art history knowledge, and individuals with low prior knowledge tended to gain more from pre- to post test. Ultimately, researchers should approach test equating based on their theoretical stance on Classical Test Theory and IRT and the respective assumptions. Regardless of the approach or method, test equating requires a representative sample of sufficient size. With small sample sizes, the application of a range of equating approaches can expose item and test features for review, inform interpretation, and identify paths for improving instruments for future study.

***Keywords*—**Effectiveness, equipercentile equating, IRT, learning games, linear equating, mean-sigma equating.

Steven W. Carruthers is now in the PhD program in Educational Psychology (Educational Technology) in the Department of Learning Sciences, Texas A&M University, College Station, TX, 77840, USA He was with the LIVElab game studio, Department of Visualization, Texas A&M University, and the Triseum learning game studio, Bryan, TX, who partnered on development of ARTé: Mecenas and supported this research (e-mail: carruthers101@gmail.com).

I. Introduction

THE purpose of the present research is to equate two test forms for pre-and post-testing as part of a study to evaluate the educational effectiveness of the ARTé: Mecenas learning game for art history. In order to compare results from the two forms, Form R and Form Q, the tests must be parallel for content and difficulty, and the scales equated so that scores from either form can be used to calculate ability and gains from pre- to post-testing across research groups. Test equating is a challenge, especially with relatively small samples. Common procedures such as mean-sigma anchor item equating in IRT, equipercentile equating, and linear equating have different practical and conceptual strengths and weaknesses, and results differ empirically and in interpretation.

II. Methods

A. Item Development

The first challenge of equating is creating parallel forms for pre- and post-testing. Test forms must be parallel for content and statistical parameters, or the scores cannot be equated by any method [1]. For the present study, test items with parallel content, item type and format, and intended difficulty were composed. A subject matter expert in art history and instructional designer composed the items and aligned them to the four learning objectives and their subtopics, on which ARTé: Mecenas was designed. The item pool consists of six anchor items (i.e., common items) on a range of content and difficulty that appear on both forms for the post-tests and two sets of 17 parallel items (core items), which address all four learning objectives.

B. Design

The study follows a quasi-experimental design with purposeful sampling of students enrolled in a college level art survey course and a counterbalancing design to distribute both test forms on the pre- and post-tests (n = 134). Random group

design ensures that participants are randomly assigned to test forms [1], and research groups, thus forming *equivalent* groups distributed across the two forms. Counterbalancing ensures that participants who take Form Q on the pre-test take Form R on the post-test (test order QR) and vice versa (test order RQ). The *equity* property states that true scores and converted scores must have the same statistical properties—mean, standard deviation, and shape—so the researcher must examine item and/or test parameters from performance data [2].

C. Procedures

In the present study, all participants gave consent and completed a profile survey. The Experimental Group participants were asked to take a pre-test, play ARTé: Mecenas online and accrue at least 4 hours of gameplay or complete Level 4 of the game within a two-week period, and take a post-test. Over the same period, the Control Group took the pre-test and a post-test, but did not play the game or have a comparable intervention during the study period. The local Institutional Review Board approved the research protocol.

Because the game content is complementary to typical course content, the researcher expected that participants would have low prior knowledge, too low to allow test equating or item scaling from pre-test results. To capture results from a representative sample including individuals with higher content knowledge, the anchor items are presented on the post-test to all participants, and post-test results are used for test scaling.

Three methods of test equating were used. IRT analysis was generated by IRTPRO software, Version 2.1 for Windows, Copyright © 2017. Additional analyses were generated using SAS University Edition software, Version 3.6 of the SAS System for macOS, Copyright© 2012-2016. The original plan was to use IRT Unequal Groups with Common Items and Mean-Sigma Equating. However, the relatively small sample and fit concerns dictated that alternative raw score methods be used for comparison, Linear Equating and Equipercentile Equating. In all three methods, one test form was selected as the reference form, and scores on the "new" form were adjusted to align to the reference form scale. For the present study, the reference form was Form Q, and Form R scores are adjusted to place them on the Form Q scale.

The method that produces optimal test parameters across forms is used for further analyses, such as measuring gain from pre- to post-test and comparing results from the Experimental Group to the Control Group. The preferred method would align forms by mean, standard deviation, and shape, and minimize any effect from the order of test forms presented to the individual in the counterbalanced design. ANOVA and ANCOVA are used to examine equating results.

IRT procedures enable the researcher to examine the item parameters and select an optimal set of items from each form to establish the difficulty and discrimination of each item and measure each individual's ability on a common scale [3]. Ability and item difficulty are expressed on the *theta* scale values that typically range from -3.0 to +3.0 (but are infinitely positive or negative). Although equivalent groups are sufficient for raw score equating methods, IRT approaches require either the same persons test on all items, which are scaled to a common scale, or two equivalent groups test on common items to align the items on different forms. The present study uses common items on each form and assumes non-equivalent groups despite randomization. The anchor items serve as an *internal anchor test*, distributed among the core test items [4]. Only common items are used to calculate the mean-sigma parameters [5]. The groups are randomly assigned, similar, and representative of a range of ability, which should aid in meeting statistical assumptions.

Mean-sigma item scaling procedures produces a slope adjustment (α) and mean (β) adjustment from common item parameters [3]. Based on a slope-intercept function, the formula $y=\alpha x+\beta$ scales an IRT difficulty parameter on Form R to scale difficulty to the reference Form Q scale. To scale the discrimination parameter, divide the item discrimination parameter by α. Those revised item parameter adjustments can be used to score persons on Form R to the Form Q scale. Likewise, an individual's Form R theta score, which is on the same scale as item difficulty, can be imputed for x in $y=\alpha x+\beta$ to calculate the score on the Form Q scale.

Reference [4] summarized several issues with conventional and IRT equating and scaling approaches. Foremost, a sufficient number of common items are required. One rule of thumb for scaling in IRT is to have 20% to 30% common items that are representative of the core test [1], and if retained, the ratio 6 common to 17 core items on the forms meets that requirement (26% common items). Concurrent scaling items may require 15 or more common items on a 40-item test with rectangular or normal information functions [4]. Also, the sample size of 134 participants in the present study is lower than conventional IRT samples of 250, 500, or more. Therefore, the present sample and test design may not suffice for accurate IRT mean-sigma equating. In addition, IRT works best in situations where the assumptions and testing parameters such as sufficient and representative sampling are met, which is a challenge.

Equipercentile equating examines raw score distribution in terms of frequency of each possible score, with the goal of having the same proportion of scores in each percentile segment (e.g., by quartile, quintile, decile) by matching scores on the new form to the score on the reference form with the same percentile rank [1]. It assumes that the proportion of individual scores at or below an equated score will be the same for either form [7]. With smaller sample sizes, however, equipercentile equating would not be as replicable as it would be with larger sample sizes [4]. Also, because a small sample size can produce rather jagged distribution, reference [4] recommended smoothing the distribution. Therefore, the researcher will apply loglinear smoothing (using SAS PROC GENMOD procedures) to compute adjusted smoothed frequencies for each possible score. By examining the distribution and percentile thresholds, the researcher adjusts scores to fit the distributions and align each new form score (Form R) to a reference form score (Form Q). Quintile

thresholds are used in the present study because deciles were too granular for a 20-point scale.

Linear equating, similar to mean-sigma equating in IRT, produces a slope and mean adjustment, which allows difficulty to vary across score level and adjusts the score on a new form (which receives the adjustment) to the mean and standard deviation on the reference form. Given the adjustment function y=ax+b, x is the raw score on Form R and y is the adjusted score on the Form Q scale. Scaling by adjusting not only mean (b) but also slope (a) can compensate for differing magnitudes of effect on lower or higher scores, depending on the slope. Linear equating does not require anchor items. Random assignment is sufficient for equivalent groups.

III. Results

The Experimental Group (n=82) completed study activities, including pre-test, gameplay, and post-test during the study period. A subset of 37 individuals completed Level 4. Over the same period, the Control Group (n=52) completed the pre- and post-tests, with no gameplay or comparable intervention. The researcher applied IRT with mean-sigma equating (scaling) and equipercentile equating and linear equating to raw scores on Form Q and Form R and compared the ANOVA eta^2 and Cohen's d effect sizes. The effect thresholds (small>0.2, medium>0.5, large>0.8) are somewhat arbitrary and do not mean that an effect is clinically or practically significant, but Cohen's d values can indicate relative differences in the same context.

A. Raw Score Analysis

For a baseline, raw scores were examined. Both forms displayed low reliability (Cronbach's alpha Form Q α=0.324 and Form R α=0.534), perhaps due to the small sample size and item discrimination. The researcher conducted a Two-Way ANOVA with raw post-test scores by group and test form, which resulted in a 4.04% eta^2 between groups effect size (df =133, p=0.0195) and a 0.52 (small) Cohen's d effect size for the post-test performance of the Control and Experimental Groups. The Group by Form mixed effect was less than 1% and not statistically significant. Form had a direct effect of 1.29% (p=0.1835). Although not statistically significant, the 1.29% effect depending on which form the individual took on the test may be enough to obscure group differences.

B. IRT with Mean-Sigma Equating

IRT with mean-sigma equating was applied to a subset of items that generated the best fit for a one-facet model on art history knowledge. The model fit of both Form Q and Form R indicated "good" fit (Q RMSEA=0.02, R RMSEA=0.05). However, only four anchor items could be retained for scaling, and the Test Information Functions peaked at 4 and 6, respectively, which indicates low score reliability. With the omitted core and anchor items, rather than a mini version of the full test forms, the smaller item pool creates a *midi* version of the test with a narrow range of difficulty [8]. Further analysis and outcomes should be interpreted with caution.

Applied to the study data, mean-sigma equating of the anchor items produced a more acute than expected ability-difficulty adjustments (α=0.6911, β=0.31), which are used to scale the new form items or ability scores to the reference scale. Individual responses were scored on the item parameters for the retained core items, and the adjustment was applied to the individual theta scores on Form R and individuals' pre- and post-tests were scored. The researcher conducted a Two-Way ANOVA with the mean-sigma adjusted scores, which resulted in a 5.41% eta^2 effect size (df=133, p=0.0049) and a 0.80 (large) Cohen's d effect size. The Group by Form mixed effect was less than 1% and not statistically significant. Form had a direct effect of 8.39% (p<0.001), which is both statistically significant and practically significant.

C. Equipercentile Equating

TABLE I
FORM R RAW TO R* SCALED SCORE CONVERSION

Quintile	Form Q	Form R	Scaled R*
	3	2	**3**
	4	3	**4**
	5	4	**5**
	6	5	**6**
1st Quintile	7	6	**7**
		7	**7.5**
2nd Quintile	8	8	8
Median	9	9	9
3rd Quintile	10	9	9
	10	10	10
4th Quintile	11	11	11
	12	12	12
	13	13	13
	14	14	14
	15	15	15
		16	16
Median	9.00	9.0	9.0
Mean	9.15	8.6	8.9
SD	2.62	3.0	2.6
Skewness	0.0986	0.233	0.471
Kurtosis	-0.743	-0.779	-0.522

Equipercentile equating was conducted on the same raw data, but with smoothed distributions calculated by using SAS PROC GENMOD procedures. The researcher examined the resulting redistributions at the quintile thresholds (20th percentiles) and adjusted Form R scores to fit Form Q distributions. Table I presents the raw and scaled Form R scores, R*, which shows lower scores from 2 to 7 required scaling. Adding 1 or 0.5 points to a few items on Form R was sufficient for aligning the medians and quintiles to equate the forms. However, with a Form R* score of 7.5 aligned to an integer on Form Y scale, this scale violates the equating principle of *symmetry*, where any equating strategy works in either direction [2]. Note that scaling maintained the median at 9 and aligned the Form R* mean and standard deviation to Form Q parameters. However, although equipercentile redistributes scores, in the present case, Form R* skewness and kurtosis (distribution at the tails) worsened compared to raw Form R scores. With minor adjustments to the raw scores,

it is not readily evident if equipercentile equating will improve alignment of Form Q and Form R.

Based on Form Q and equipercentile scaled Form R* scores, the researcher conducted a Two-Way ANOVA between the Experimental Group and Control Group with the equipercentile adjusted scores, which resulted in a 4.07% effect size (df=133, p=0.0179) and a 0.64 (medium) Cohen's d effect size. The Group by Form mixed effect was less than 1% and not statistically significant. However, Form had a 3.18% direct effect (p=0.0361). Ultimately, equipercentile equating worsened test shape compared to the raw score results, despite better alignment of the test means and standard deviations.

D.Linear Equating

Linear equating produced the slope and mean adjustment variables (a=0.66565, b=0.31), and the function used to equate scores from all core items on Form R to the form Q scale. The researcher conducted a Two-Way ANOVA between the Experimental Group and Control Group with the linear equating adjusted scores, which resulted in a between group effect size of 5.6% (n=134, df=133, p=0.0240) and a 0.70 (medium) Cohen's d effect size between the post-test scores by the Control and Experimental Groups. The Group and Form mixed effect was less than 1% and not statistically significant. Form had no direct effect, approaching 0.0% (p=0.9560), because the scaling parameters aligned Form R* mean and standard deviation to the Form Q scale parameters.

IV. Game Effectiveness

Based on Form Q and linearly equated and adjusted Form R* scores, a full-factorial Two-Way ANOVA was conducted to examine the effect size, with Gain (shift) from pre- to post-test as the dependent variable by research group and test form. Table II presents the ANOVA table with eta^2 effect sizes. The between group effect size for Experimental Group (n=82, mean=3.78056, SD=2.88412) and Control Group (n=52, mean=2.43446, SD=2.39549) was 5.46% (p<0.001) and Cohen's d effect size of 0.51 (medium) with less than 1% effect size attributed to test form or the mixed effect, which were not statistically significant.

An ANOVA was conducted on the subset of 37 Experimental Group participants who completed Level 4 of the game. Table III presents the ANOVA table with eta^2 effect sizes. The between group effect size for Experimental Group who completed the game (n=37, mean=4.32968, SD=2.93066) and Control Group (n=52, 2.43447, SD=2.39549) was 11.3% (p=0.0013) and Cohen's d effect size of 0.71 (medium) with less than 1% effect attributed to test form and 1.4% mixed effect, which were not statistically significant. The 1.4% mixed effect may relate to the Control Group having lower scores on the post-test, in the range of greater score adjustment that was identified in the equipercentile equating of Form R*.

Independent ANOVAs were used to calculate effect sizes from pre- to post-test. The Control Group gained 26.59% (p<0.001), the Experimental Group who played ARTé: Mecenas gained 37.80% (p<0.001), and the Experimental Group who completed the game gained 49.71% (p<0.001) on average. The Experimental Group who experienced the game intervention gained more in art history knowledge compared to the Control Group, and the effect was greater for those who completed the game. The measurable gains by the Control Group, who did not experience the game intervention, may be attributed to continued art history instruction as part of their course and/or familiarity with the test format. However, counterbalancing ensured that participants saw different items on the pre- and post-tests.

TABLE II
Two-Way ANOVA on Gain by Form and Group

Source	DF	SS	Mean Square	F Value	Pr > F	eta^2
Model	3	66.8796	22.2932	3.03	0.0319	6.53%
FORM	1	8.7562	8.7562	1.19	0.2775	0.86%
GROUP	1	55.8649	55.8649	7.59	0.0067	5.46%
FORM*GROUP	1	2.2586	2.2586	0.31	0.5806	0.22%
Error	130	957.2097	7.3632			
Corrected Total	133	1024.0893				

Pre-to post-test gain as dependent variable by research group and form using linearly equated scores.

TABLE III
Two-Way ANOVA on Gain by Form and Group for for Control vs. Experimental Group Participants Who Completed the Game

Source	DF	SS	Mean Square	F Value	Pr > F	eta^2
Model	3	87.5198	29.1733	4.19	0.0081	12.90%
FORM	1	1.2929	1.2929	0.19	0.6674	0.19%
GROUP	1	76.7213	76.7213	11.03	0.0013	11.30%
FORM*GROUP	1	9.5056	9.5056	1.37	0.2456	1.40%
Error	85	591.1669	6.9549			
Corrected Total	88	678.6867				

Pre-to post-test gain as dependent variable by research group and form using linearly equated scores.

To examine the effect size in terms of prior knowledge, a full-factorial ANCOVA was conducted. Post-Test scores are the dependent variable, regressed on by research group with pre-test scores as covariate. Table IV presents the ANCOVA table with eta^2 effect sizes. The between group effect size for Experimental Group (n=82) and Control Group (n=52) was

9.94% (p<0.001) with a 3.84% (p=0.0176) effect size attributed to pre-test score and mixed effect less than 1% and not statistically significant. Pre-test scores correlate with post-test scores (Pearson r=0.325) but also negatively correlate with gain (Pearson r=-0.358). In short, individuals with low pre-test scores tended to gain more from pre- to post test. The full model effect size was 13.83% (p < 0.001).

The between group effect size for Experimental Group participants who completed Level 4 (n=37) and Control Group (n=52) was 14.39% (p<0.001) with a 4.77% (p=0.0274) effect size attributed to pre-test score. Table V presents the ANCOVA table. The full model effect size was 19.55% (p<0.001).

TABLE IV
ANCOVA ON POST-TEST BY GROUP WITH PRE-TEST SCORE COVARIATE

Source	DF	SS	Mean Square	F Value	Pr > F	eta^2
Model	3	131.8586	43.9529	6.95	0.0002	13.83%
GROUP	1	94.7989	94.7989	14.99	0.0002	9.94%
PRETEST	1	36.5868	36.5868	5.79	0.0176	3.84%
PRETEST*GROUP	1	0.4729	0.4729	0.07	0.7849	0.05%
Error	130	821.8958	6.3223			
Corrected Total	133	953.7544				

Post-test scores as dependent variable regressed on by research group with pre-test covariate using linearly equated scores.

TABLE V
ANCOVA ON POST-TEST BY GROUP WITH PRE-TEST SCORE COVARIATE FOR CONTROL VS. EXPERIMENTAL GROUP PARTICIPANTS WHO COMPLETED THE GAME

Source	DF	SS	Mean Square	F Value	Pr > F	eta^2
Model	3	117.2583	39.0861	6.88	0.0003	19.55%
GROUP	1	86.2611	86.2611	15.19	0.0002	14.38%
PRETEST	1	28.6054	28.6054	5.04	0.0274	4.77%
PRETEST*GROUP	1	2.3918	2.3918	0.42	0.5180	0.40%
Error	85	482.5547	5.6771			
Corrected Total	88	599.8130				

Post-test scores as dependent variable regressed on by research group with pre-test covariate using linearly equated scores.

V. DISCUSSION AND CONCLUSION

By applying more than one method of item scaling and form equating, the researcher illustrated issues that arise in test equating with smaller sample sizes. In the raw score analysis, both forms displayed low reliability (Cronbach's alpha Form Q α =0.324 and Form R α=0.534), perhaps due to the small sample size and low item discrimination. Based on the ANOVA analysis of post-test results by Group (Experimental and Control), Form had a direct effect of 1.29% (p=0.1835). Although not statistically significant, the 1.29% effect depending on which form the individual took on the test may be enough to obscure group differences. With the caveat that improved score reliability is required, a well-constructed test may still perform adequately as a whole and enable the researcher to rely on raw scores.

The intended equating approach was IRT by mean-sigma equating of anchor items parameters. With the relatively small sample size (n=134), IRT procedures helped identify items that did not discriminate well and refine the assessment through omitting those items. Omitting items decreased the scope of content the assessment covers, which affects interpretation of the test results, and reduced the number of common anchor items from 6 to 4.

The mean-sigma parameters derived from those anchor items produced a more acute adjustment that resulted in worsening the influence of Form on results. Form had a direct effect of 8.39% (p<0.001), which is both statistically significant and practically significant. The scaling based on just four anchor items was not sufficiently robust to produce accurate mean-sigma scaling parameters. A revised test should include additional common items to refine mean-sigma parameters, or the research should scale the items by single group design [3]. Ultimately, all items should be reexamined with data from a larger sample size to validate test performance and item scaling.

Equipercentile equating did not perform well in the present case, but the procedure did help highlight the lower end of the score range as a source of misalignment of the forms. Based on this evidence, revision of the forms and items could start with reviewing test performance at lower ability levels.

In the present case, linear equating performed best out of the three equating methods with the relatively small sample size. It attenuated Form effect statistically to approach 0.0% (p=0.9560) by aligning Form R* to the Form Q scale through a slope and mean adjustment. Some items may have performance issues, which are not assessed by linear equating.

Based on an ANOVA using the linearly equated scores, the Experimental (n=82) and Control (n=52) between group effect size is 5.46% (p <0.001), and for the subset of Experimental Group participants (n = 37), the effect size rises to 11.3% (p=0.0013). In a parallel ANCOVA to account for the effect of prior knowledge, the between group effect size for Experimental Group (n=82) and Control Group was 9.94% (p<0.001) with a 3.84% (p=0.0176) effect of prior knowledge (as measured by the pre-test). In short, individuals with lower pre-test scores (lower prior knowledge) tended to gain more from pre- to post test. For the subset of Experimental Group participants who completed the game, the between group

effect rises to 14.38% ($p<0.001$) with a prior knowledge covariate effect of 4.77% ($p=0.0274$). Therefore, the current research indicates that playing and completing the ARTé: Mecenas game increased art history knowledge, and on average the effect is greater for individuals with lower levels of prior knowledge. However, raw score reliability is low, so future research should use a larger sample size to reassess the scale and equating. In addition, future research should add a comparative learning intervention for an experimental design that examines the effect of gameplay to alternative instruction such as interactive video lessons.

Ultimately, researchers should approach test equating based on their theoretical stance, such as their knowledge of Classical Test Theory and IRT and the respective assumptions. Reference [6] suggested it is better to use the simpler, more easily rationalized and explained method with realistic underlying assumptions. In the present case, linear equating best met those requirements. Regardless of the approach or method, test equating requires a representative sample of sufficient size. With small sample sizes, the application of a range of equating approaches can help expose item and test features for review, inform interpretation, and identify paths for improving instruments for future study.

ACKNOWLEDGMENT

S. W. Carruthers thanks André Thomas, Director of the LIVElab game studio, CEO of the Triseum learning game studio, and Texas A&M University faculty primary investigator on the IRB protocol; the respective game development teams; and Dr. Livia Stoenescu, Department of Visualization, Texas A&M University, subject matter expert on the art history assessment used in this study.

REFERENCES

[1] M. J. Kolen and R. L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices.* New York: Springer, 2014.
[2] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum, 1980.
[3] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*, vol. 2, New York: Sage, 1991.
[4] L. L. Cook and N. S. Paterson, "Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances," *Applied Psychological Measurement*, vol. 11, no. 3, pp. 225-244, 1987.
[5] J. González, "SNSequate: Standard and nonstandard statistical models and methods for test equating," *Journal of Statistical Software*, vol. 59, no. 7, pp. 1-30, 2014.
[6] S. A. Livingston, "Equating Test Scores (Without IRT)." Princeton, NJ: *Educational Testing Service*, 2004.
[7] M. J. Kolen, "Effectiveness of analytic smoothing in equipercentile equating," *Journal of Educational Statistics*, vol. 9, no. 1, pp. 25-44, 1984.
[8] J. P. Meyer and S. Zhu, "Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating," *Research & Practice in Assessment*, vol. 8, 2013.