RECENT SCALING, LINKING, & EQUATING RELATED PUBLICATIONS

2010–2022

NOVEMBER 2022

NCME Special Interest Groups in Measurement in Education (SIGIMIE)
Contemporary Issues in Scaling, Linking, & Equating (SLE)

ABOUT US

The Contemporary Issues in Scaling, Linking, and Equating (SLE) SIGIMIE focuses on newly evolving measurement issues related scaling, linking and equating. We encourage scholarly development by connecting professionals from diverse backgrounds to discuss and take action on contemporary issues related to SLE. We are sponsored by the National Council on Measurement in Education (NCME).

ABOUT THIS SUMMARY

This summary provides an evolving list of recent publications related to SLE. Published articles are sourced from a collection of journals including *Applied Measurement in Education* (AME), *Applied Psychological Measurement* (APM), *Educational Measurement: Issues and Practices* (EM:IP), *Educational and Psychological Measurement* (EPM), *Journal of Educational and Behavioral Statistics* (JEBS), *Journal of Educational Measurement* (JEM), and *Psychometrika*. Some published research reports with the focus on SLE methods and applications are also included. Feel free to email us (*sle.sigimie@gmail.com*) with any recent publications that are not listed here so that we may review and add them. Despite our best efforts, typing mistakes are likely to exist; please let us know if you notice any of them.

SLE SIGIMIE EXECUTIVE COMMITTEE (2022–23)

- Chair: Mengyao Zhang
- Vice Chair: Kyung Yong Kim
- Secretary: S. Kanageswari Suppiah Shanmugam
- Graduate Student Liaison: Tong Wu

TABLE OF CONTENTS

Applied Measurement in Education (AME): 2010–2022

Title	The Effect of Repeaters on Equating
Abstract	Test equating might be affected by including in the equating analyses examinees who have taken the test previously. This study evaluated the effect of including such repeaters on Medical College Admission Test (MCAT) equating using a population invariance approach. Three-parameter logistic (3-PL) item response theory (IRT) true score and traditional equipercentile equating methods were used under the random groups equating design. This study also examined whether or not population sensitivity of equating by repeater status varies depending on other background variables (gender and ethnicity). The results indicated that there was some evidence of repeaters' effect on equating with varying amounts of such effect by gender.
Citation	Kim, H., & Kolen, M. J. (2010). The effect of repeaters on equating. *Applied Measurement in Education*, *23*(3), 242–265. https://doi.org/10.1080/08957347.2010.486024
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2010.486024

Title	Practical Application of a Synthetic Linking Function on Small-Sample Equating
Abstract	The synthetic function is a weighted average of the identity (the linking function for forms that are known to be completely parallel) and a traditional equating method. The purpose of the present study was to investigate the benefits of the synthetic function on small-sample equating using various real data sets gathered from different administrations of tests from a licensure testing program. We investigated the chained linear, Tucker, Levine, and mean equating methods, along with the identity and the synthetic functions with small samples (N = 19 to 70). The synthetic function did not perform as well as did other linear equating methods because test forms differed markedly in difficulty; thus, the use of the identity function produced substantial bias. The effectiveness of the synthetic function depended on the forms' similarity in difficulty.
Citation	Kim, S., von Davier, A. A., & Haberman, S. (2011). Practical application of a synthetic linking function on small-sample equating. *Applied Measurement in Education*, *24*(2), 95–114. https://doi.org/10.1080/08957347.2011.554601
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2011.554601

Title	Collateral Information for Equating in Small Samples: A Preliminary Investigation
Abstract	This article describes a preliminary investigation of an empirical Bayes (EB) procedure for using collateral information to improve equating of scores on test forms taken by small numbers of examinees. Resampling studies were done on two different forms of the same test. In each study, EB and non-EB versions of two equating methods—chained linear and chained mean—were applied to repeated small samples drawn from a large data set collected for a common-item equating. The criterion equating was the chained linear equating in the large data set. Equatings of other forms of the same test provided the collateral information. New-form sample size was varied from 10 to 200; reference-form sample size was constant at 200. One of the two new forms did not differ greatly in difficulty from its reference form, as was the case for the equatings used as collateral information. For this form, the EB procedure improved the accuracy of equating with new-form samples of 50 or fewer. The other new form was much more difficult than its reference form; for this form, the EB procedure made the equating less accurate.
Citation	Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating in small samples: A preliminary investigation. *Applied Measurement in Education*, *24*(4), 302–323. https://doi.org/10.1080/08957347.2011.607057
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2011.607057

Title	Investigating Repeater Effects on Chained Equipercentile Equating with Common Anchor Items
Abstract	This study investigated the impact of repeat takers of a licensure test on the equating functions in the context of a nonequivalent groups with anchor test (NEAT) design. Examinees who had taken a new, to-be-equated form of the test were divided into three subgroups according to their previous testing experience: (a) repeaters who previously took the reference form, to which the new form would be equated; (b) repeaters who previously took any form other than the reference form; and (c) first-time test-takers for whom the new form was the first exposure to the test. Equating functions remained essentially invariant across all repeaters versus first-time test-takers, supporting score equatability of the two forms. However, when the repeater subgroup was sub-divided based on the particular form examinees took previously, subgroup equating functions substantially differed from the total-group equating function, indicating subgroup dependency of score equating. The results indicate that repeater membership needs to be more clearly specified to assess the impact of repeaters on score equating. Such clarification may be especially necessary for high-stakes licensure tests because repeaters tend to perform more poorly on such tests than first-time test-takers.
Citation	Kim, S. & Walker, M. E. (2012). Investigating repeater effects on chained equipercentile equating with common anchor items. *Applied Measurement in Education*, *25*(1), 41–57. https://doi.org/10.1080/08957347.2012.635481
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2012.635481

Title	The Impact of Multidirectional Item Parameter Drift on IRT Scaling Coefficients and Proficiency Estimates
Abstract	Item parameter drift (IPD) occurs when item parameter values change from their original value over time. IPD may pose a serious threat to the fairness and validity of test score interpretations, especially when the goal of the assessment is to measure growth or improvement. In this study, we examined the effect of multidirectional IPD (i.e., some items become harder while other items become easier) on the linking procedure and rescaled proficiency estimates. The impact of different combinations of linking items with various multidirectional IPD on the test equating procedure was investigated for three scaling methods (mean-mean, mean-sigma, and TCC method) via a series of simulation studies. It was observed that multidirectional IPD had a substantive effect on examinees' scores and achievement level classifications under some of the studied conditions. Choice of linking method had a direct effect on the results, as did the pattern of IPD.
Citation	Han, K. T., Wells, C. S., & Sireci, S. G. (2012). The impact of multidirectional item parameter drift on IRT scaling coefficients and proficiency estimates. *Applied Measurement in Education*, *25*(2), 97–117. https://doi.org/10.1080/08957347.2012.660000
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2012.660000

Title	Determining the Anchor Composition for a Mixed-Format Test: Evaluation of Subpopulation Invariance of Linking Functions
Abstract	This study examined the appropriateness of the anchor composition in a mixed-format test, which includes both multiple-choice (MC) and constructed-response (CR) items, using subpopulation invariance indices. Linking functions were derived in the nonequivalent groups with anchor test (NEAT) design using two types of anchor sets: (a) MC only and (b) a mix of MC and CR. In each anchor condition, the linking functions were also derived separately for males and females, and those subpopulation functions were compared to the total group function. In the MC-only condition, the difference between the subpopulation functions and the total group function was not trivial in a score region that included cut scores, leading to inconsistent pass/fail decisions for low-performing examinees in particular. Overall, the mixed anchor was a better choice than the MC-only anchor to achieve subpopulation invariance between males and females. The research reinforces subpopulation invariance indices as a means of determining the adequacy of the anchor.
Citation	Kim, S., & Walker, M. (2012). Determining the anchor composition for a mixed-format test: Evaluation of subpopulation invariance of linking functions. Applied *Measurement in Education*, *25*(2), 178–195. https://doi.org/10.1080/08957347.2010.524720
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2010.524720

Title	Considerations for Equating Alternate Assessments: Two Case Studies of Alternate Assessments Based on Alternate Achievement Standards
Abstract	The development of alternate assessments for students with disabilities plays a pivotal role in state and national accountability systems. An important assumption in the use of alternate assessments in these accountability systems is that scores are comparable on different test forms across diverse groups of students over time. The use of test equating is a common way that states attempt to establish score comparability on different test forms. However, equating presents many unique, practical, and technical challenges for alternate assessments. This article provides case studies of equating for two alternate assessments in Michigan and an approach to determine whether or not equating would be preferred to not equating on these assessments. This approach is based on examining equated score and performance-level differences and investigating population invariance across subgroups of students with disabilities. Results suggest that using an equating method with these data appeared to have a minimal impact on proficiency classifications. The population invariance assumption was suspect for some subgroups and equating methods with some large potential differences observed.
Citation	Wyse, A. E., Dean, V. J., Viger, S. G., & Vansickle, T. R. (2013). Considerations for equating alternative assessments: Two case studies of alternative assessments based on alternate achievement standards. *Applied Measurement in Education*, *26*(1), 50–72. https://doi.org/10.1080/08957347.2013.739460
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2013.739460

Title	Selection of Common Items as an Unrecognized Source of Variability in Test Equating: A Bootstrap Approximation Assuming Random Sampling of Common Items
Abstract	The standard error of equating quantifies the variability in the estimation of an equating function. Because common items for deriving equated scores are treated as fixed, the only source of variability typically considered arises from the estimation of common-item parameters from responses of samples of examinees. Use of alternative, equally appropriate anchor sets results in different equating transformations, even when standard errors of common-item parameter estimates are negligible. A bootstrap approximation for quantifying the variability due to common-item selection is derived for a statewide assessment assuming random selection of common items from a large hypothetical item pool. The standard error of equating and the error due to common-item selection constitute only a small fraction of the variability for individual examinee scores. For group means, other sources of error shrink as sample size increases; unaffected by sample size, error due to common-item selection may become the dominant source of variability.
Citation	Michaelides, M. P. (2013). Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, *27*(1), 46–57. https://doi.org/10.1080/08957347.2013.853069
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2013.853069

Title	Impact of Accumulated Error on Item Response Theory Pre-Equating With Mixed Format Tests
Abstract	The equating of tests is an essential processes in high-stakes, large-scale testing conducted over multiple forms or administrations. By adjusting for differences in difficulty and placing scores from different administrations of a test on a common scale, equating allows scores from these different forms and administrations to be directly compared to one another (Kolen & Brennan, 2004); as such, the importance of the accuracy of equating is paramount for any assessment program. Due to the increasingly fast turnaround times required of testing companies in the reporting of test scores, many testing programs rely on pre-equating methodologies. Although many testing companies report the use of pre-equating for some of their tests, there is very little research evidence supporting its use. This study seeks to determine the impact of accumulated error from the long-term use of pre-equating in an operational testing program. Results indicate that for some tests, the pre-equating and post-equating produced almost identical results, while in other cases the differences were quite dramatic.
Citation	Keller, L. A., Keller, R., Cook, R. J., & Colvin, K. F. (2015). Impact of accumulated error on item response theory pre-equating with mixed format tests. *Applied Measurement in Education*, *29*(1), 65–82. https://doi.org/10.1080/08957347.2015.1102912
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2015.1102912

Title	An Extension of IRT-Based Equating to the Dichotomous Testlet Response Theory Model
Abstract	Current procedures for equating number-correct scores using traditional item response theory (IRT) methods assume local independence. However, when tests are constructed using testlets, one concern is the violation of the local item independence assumption. The testlet response theory (TRT) model is one way to accommodate local item dependence. This study proposes methods to extend IRT true score and observed score equating methods to the dichotomous TRT model. We also examine the impact of local item dependence on equating number-correct scores when a traditional IRT model is applied. Results of the study indicate that when local item dependence is at a low level, using the three-parameter logistic model does not substantially affect number-correct equating. However, when local item dependence is at a moderate or high level, using the three-parameter logistic model generates larger equating bias and standard errors of equating compared to the TRT model. However, observed score equating is more robust to the violation of the local item independence assumption than is true score equating.
Citation	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. https://doi.org/10.1080/08957347.2016.1138956
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2016.1138956

Title	Bi-Factor MIRT Observed-Score Equating for Mixed-Format Tests
Abstract	The main purposes of this study were to develop bi-factor multidimensional item response theory (BF-MIRT) observed-score equating procedures for mixed-format tests and to investigate relative appropriateness of the proposed procedures. Using data from a large-scale testing program, three types of pseudo data sets were formulated: matched samples, pseudo forms, and simulated data sets. Very minor within-format residual dependence in mixed-format tests was found after controlling for the influence of the primary general factor. The unidimensional IRT and BF-MIRT equating methods produced similar equating results for the data used in this study. When a BF-MIRT model is implemented, we recommend the use of observed-score equating instead of true-score equating because the latter requires an arbitrary approximation or reduction process to relate true scores on test forms.
Citation	Lee, G., & Lee, W.-C. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education*, *29*(3), 224–241. https://doi.org/10.1080/08957347.2016.1171770
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2016.1171770

Title	IRT Item Parameter Scaling for Developing New Item Pools
Abstract	Increasing use of item pools in large-scale educational assessments calls for an appropriate scaling procedure to achieve a common metric among field-tested items. The present study examines scaling procedures for developing a new item pool under a spiraled block linking design. The three scaling procedures are considered: (a) concurrent calibration, (b) separate calibration with one linking, and (c) separate calibration with three sequential linking. Evaluation across varying sample sizes and item pool sizes suggests that calibrating an item pool simultaneously results in the most stable scaling. The separate calibration with linking procedures produced larger scaling errors as the number of linking steps increased. The Haebara's item characteristic curve linking resulted in better performances than the test characteristic curve (TCC) linking method. The present article provides an analytic illustration that the test characteristic curve method may fail to find global solutions in polytomous items. Finally, comparison of the single- and mixed-format item pools suggests that the use of polytomous items as the anchor can improve the overall scaling accuracy of the item pools.
Citation	Kang, H., Lu, Y., & Chang, H. (2017). IRT item parameter scaling for developing new item pools. *Applied Measurement in Education*, *30*(1), 1–15. https://doi.org/10.1080/08957347.2016.1243537
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2016.1243537

Title	Application of IRT Fixed Parameter Calibration to Multiple-Group Test Data
Abstract	In applications of item response theory (IRT), fixed parameter calibration (FPC) has been used to estimate the item parameters of a new test form on the existing ability scale of an item pool. The present paper presents an application of FPC to multiple examinee groups test data that are linked to the item pool via anchor items, and investigates the performance of FPC relative to an alternative approach, namely independent 0–1 calibration and scale linking. Two designs for linking to the pool are proposed that involve multiple groups and test forms, for which multiple-group FPC can be effectively used. A real-data study shows that the multiple-group FPC method performs similarly to the alternative method in estimating ability distributions and new item parameters on the scale of the item pool. In addition, a simulation study shows that the multiple-group FPC method performs nearly equally to or better than the alternative method in recovering the underlying ability distributions and the new item parameters.
Citation	Kim, S., & Kolen, M. J. (2019). Application of IRT fixed parameter calibration to multiple-group test data. *Applied Measurement in Education*, *32*(4), 310–324. https://doi.org/10.1080/08957347.2019.1660344
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2019.1660344

Title	Some Methods and Evaluation for Linking and Equating with Small Samples
Abstract	The purpose of the current article is to introduce the equating and evaluation methods used in this special issue. Although a comprehensive review of all existing models and methodologies would be impractical given the format, a brief introduction to some of the more popular models will be provided. A brief discussion of the conditions required for equating precedes the discussion of the equating methods themselves. The procedures in this review include the Tucker method, mean equating, nominal weights mean, simplified circle arc, identity equating, and IRT/Rasch model equating. Models shown that help to evaluate the success of the equating process are the standard error of equating, bias, and root-mean-square error. This should provide readers with a basic framework and enough background information to follow the studies found in this issue.
Citation	Peabody, M. R. (2020). Some methods and evaluation for linking and equating with small samples. *Applied Measurement in Education*, *33*(1), 3–9. https://doi.org/10.1080/08957347.2019.1674304
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2019.1674304

Title	Effect of Sample Size on Common Item Equating Using the Dichotomous Rasch Model
Abstract	This study addresses equating issues with varying sample sizes using the Rasch model by examining how sample size affects the stability of item calibrations and person ability estimates. A resampling design was used to create 9 sample size conditions (200, 100, 50, 45, 40, 35, 30, 25, and 20), each replicated 10 times. Items were recalibrated using each of these 90 samples. The deviation of these calibrations from the full sample (N = 9,678) calibrations were then computed. The ability estimates for all 9,678 examinees were then recomputed 90 times using the item calibrations from each of the 90 different samples. The deviation of 90 sets of ability estimates from the original set of ability estimates was computed. This study found that less precision and item calibration instability occur with smaller sample sizes; however, the decreasing sample size has minimal effect on the person ability estimates.
Citation	O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of sample size on common item equating using the Rasch model. *Applied Measurement in Education*, *33*(1), 10–23. https://doi.org/10.1080/08957347.2019.1674309
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2019.1674309

Title	Equating with Small and Unbalanced Samples
Abstract	Recent research has suggested that re-setting the standard for each administration of a small sample examination, in addition to the high cost, does not adequately maintain similar performance expectations year after year. Small-sample equating methods have shown promise with samples between 20 and 30. For groups that have fewer than 20 students, options are scarcer. This simulation study examined balanced and unbalanced designs across nine equating models including both classic equating models and small-sample models. The study also accounted for varying sample sizes, differences in form difficulty and candidate population, and the size of the anchor set. This study supports the use of nominal weights approaches in combination with either circle-arc or mean equating. Consistent with other research, this study found that the best ways to improve equating results are increases in sample size and/or the number of anchor items across the old and new forms. However, a testing program's tolerance for reuse will influence the decision to pool administrations.
Citation	Goodman, J., Dallas, A. D., & Fan, F. (2020). Equating with small and unbalanced samples, *Applied Measurement in Education*, *33*(1), 34–43. https://doi.org/10.1080/08957347.2019.1674311
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2019.1674311

Title	Investigating the Classification Accuracy of Rasch and Nominal Weights Mean Equating with Very Small Samples
Abstract	Maintaining equivalent performance standards across forms is a psychometric challenge exacerbated by small samples. In this study, the accuracy of two equating methods (Rasch anchored calibration and nominal weights mean) and four anchor item selection methods were investigated in the context of very small samples (N = 10). Overall, nominal weights mean equating slightly outperformed Rasch equating for three of the four anchor item selection methods, but Rasch equating slightly outperformed nominal weights mean equating when anchor items were selected to be near the cut score. The results largely confirmed previous research on the utility of nominal weights mean equating for very small samples. In addition, the results provide useful guidance for small volume programs who wish to consider using Rasch for building and equating new forms. Lastly, the results underscored the importance of being mindful about the method for selecting anchor items when building new forms.
Citation	Furter, R. T., & Dwyer, A. C. (2020). Investigating the classification accuracy of Rasch and nominal weights mean equating with very small samples. *Applied Measurement in Education*, *33*(1), 44–53. https://doi.org/10.1080/08957347.2019.1674307
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2019.1674307

Title	Investigating Repeater Effects on Small Sample Equating: Include or Exclude?
Abstract	Examinees who attempt the same test multiple times are often referred to as “repeaters.” Previous studies suggested that repeaters should be excluded from the total sample before equating because repeater groups are distinguishable from non-repeater groups. In addition, repeaters might memorize anchor items, causing item drift under a non-equivalent anchor test (NEAT) design. However, under small sample equating conditions, removing repeaters might lead to smaller sample size, which increases sampling errors. Therefore, three solutions were investigated in the current study: 1) excluding repeaters, 2) excluding drifted anchor items, and 3) applying Rasch true score equating to maintain the population invariance if repeaters were removed. The results suggested excluding repeaters if the anchor were exposed to them. Circle arc equating can be applied if it is impossible to exclude all drifted anchor items. Applying Rasch true did outperform solutions.
Citation	Dian, H., & Keller, L. (2020). Investigating repeater effects on small sample equating: Include or exclude. *Applied Measurement in Education*, *33*(1), 54–66. https://doi.org/10.1080/08957347.2019.1674302
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2019.1674302

Title	Subscore Equating and Profile Reporting
Abstract	The purpose of this study is to address the necessity of subscore equating and to evaluate the performance of various equating methods for subtests. Assuming the random groups design and number-correct scoring, this paper analyzed real data and simulated data with four study factors including test dimensionality, subtest length, form difference in difficulty, and sample size. The results indicated that reporting subscores without equating provides misleading information in terms of score profiles and that reporting subscores without a pre-specified test specification brings practical issues such as constructing alternate subtest forms with comparable difficulty, conducting equating between forms with different lengths, and deciding an appropriate score scale to be reported.
Citation	Lim, E., & Lee, W.-C. (2020). Subscore equating and profile reporting. *Applied Measurement in Education, 33*(2), 95–112. https://doi.org/10.1080/08957347.2020.1732381
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2020.1732381

Title	Asymptotic Standard Errors of Equating Coefficients Using the Characteristic Curve Methods for the Graded Response Model
Abstract	The characteristic curve methods have been applied to estimate the equating coefficients in test equating under the graded response model (GRM). However, the approaches for obtaining the standard errors for the estimates of these coefficients have not been developed and examined. In this study, the delta method was applied to derive the mathematical formulas for computing the asymptotic standard errors for the parameter scale transformation coefficients and the true score equating coefficients that are estimated using the characteristic curve methods in test equating under the GRM in the context of the common-item nonequivalent groups equating design. Simulated and real data were further used to examine the accuracy of the derivations and compare the performance of the newly developed delta method with that of the multiple imputation method. The results indicated that the standard errors produced by the delta method were extremely close to the criterion empirical standard errors as well as those yielded by the multiple imputation method. The development of the standard error expressions by the delta method in the study has important practical implications.
Citation	Zhang, Z. (2020). Asymptotic standard errors of equating coefficients using the characteristic curve methods for the graded response model. *Applied Measurement in Education*, *33*(4), 309–330. https://doi.org/10.1080/08957347.2020.1789142
Link	https://www.tandfonline.com/doi/full/10.1080/08957347.2020.1789142

Applied Psychological Measurement (APM): 2010–2022

Title	Eqboot and Eqwinboot: Java Applications for Estimating Equating Constants and the Standard Error of Equating Using IRT Methods
Abstract	(N/A)
Citation	Meyer, J. P. (2010). Eqboot and Eqwinboot: Java applications for estimating equating constants and the standard error of equating using IRT methods. *Applied Psychological Measurement*, *34*(1), 66–67. https://doi.org/10.1177/0146621609336541
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621609336541

Title	Sensitivity of Equated Aggregate Scores to the Treatment of Misbehaving Common Items
Abstract	(N/A)
Citation	Michaelides, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. *Applied Psychological Measurement*, *34*(5), 365–369. https://doi.org/10.1177/0146621609359626
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621609359626

Title	An Extension of Least Squares Estimation of IRT Linking Coefficients for the Graded Response Model
Abstract	The three types (generalized, unweighted, and weighted) of least squares methods, proposed by Ogasawara, for estimating item response theory (IRT) linking coefficients under dichotomous models are extended to the graded response model. A simulation study was conducted to confirm the accuracy of the extended formulas, and a real data study was carried out to compare the performance of the least squares methods with that of moment and characteristic curve linking methods. As found in Ogasawara's study, the generalized least squares method had the smallest asymptotic standard errors but the largest biases of linking coefficient estimates, whereas the unweighted least squares method had the largest asymptotic standard errors but the smallest biases. The weighted least squares method was intermediate. The comparison study showed that with large samples, the weighted least squares method performed nearly as well as the characteristic curve methods in linking accuracy.
Citation	Kim, S. (2010). An extension of least squares estimation of IRT linking coefficients for the graded response model. *Applied Psychological Measurement*, *34*(7), 505–520. https://doi.org/10.1177/0146621609344847
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621609344847

Title	A Comparison of Anchor-Item Designs for the Concurrent Calibration of Large Banks of Likert-Type Items
Abstract	Current interest in measuring quality of life is generating interest in the construction of computerized adaptive tests (CATs) with Likert-type items. Calibration of an item bank for use in CAT requires collecting responses to a large number of candidate items. However, the number is usually too large to administer to each subject in the calibration sample. The concurrent anchor-item design solves this problem by splitting the items into separate subtests, with some common items across subtests; then administering each subtest to a different sample; and finally running estimation algorithms once on the aggregated data array, from which a substantial number of responses are then missing. Although the use of anchor-item designs is widespread, the consequences of several configuration decisions on the accuracy of parameter estimates have never been studied in the polytomous case. The present study addresses this question by simulation, comparing the outcomes of several alternatives on the configuration of the anchor-item design. The factors defining variants of the anchor-item design are (a) subtest size, (b) balance of common and unique items per subtest, (c) characteristics of the common items, and (d) criteria for the distribution of unique items across subtests. The results of this study indicate that maximizing accuracy in item parameter recovery requires subtests of the largest possible number of items and the smallest possible number of common items; the characteristics of the common items and the criterion for distribution of unique items do not affect accuracy.
Citation	García-Pérez, M. A., Alcalá-Quintana, R., & García-Cueto, E. (2010). A comparison of anchor-item designs for the concurrent calibration of large banks of Likert-type items. *Applied Psychological Measurement*, *34*(8), 580–599. https://doi.org/10.1177/0146621609351259
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621609351259

Title	Local Observed-Score Equating with Anchor-Test Designs
Abstract	For traditional methods of observed-score equating with anchor-test designs, such as chain and poststratification equating, it is difficult to satisfy the criteria of equity and population invariance. Their equatings are therefore likely to be biased. The bias in these methods was evaluated against a simple local equating method in which the anchor-test score was used as a proxy of the proficiency measured by the test and the equating was conditional on this score. The results showed substantial bias for the two traditional methods under a variety of conditions but much smaller bias for the local method. In addition, unlike the traditional methods, the local method appeared to be quite robust with respect to changes in the difficulty and accuracy of the two tests that were equated. But like these methods, it appeared to be sensitive to a decrease in the accuracy of the anchor test as a proxy of the ability measured by the tests.
Citation	van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement*, *34*(8), 620–640. https://doi.org/10.1177/0146621609349803
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621609349803

Title	Multidimensional Linking for Domain Scores and Overall Scores for Nonequivalent Groups
Abstract	The No Child Left Behind Act requires state assessments to report not only overall scores but also domain scores. To see the information on students' overall achievement, progress, and detailed strengths and weaknesses, and thereby identify areas for improvement in educational quality, students' performances across years or across forms need to be made comparable in terms of both overall scores and domain scores. Multidimensional item response theory (MIRT) and its linking procedures for item parameter recovery have been studied; however, the effects of those linking procedures on the recovery of domain scores and overall scores have not been addressed, especially with regard to MIRT models of higher dimensions. The relationship between an examinee's overall score and domain scores can be complex and may not be adequately represented by a linear function. This article proposes using the MIRT maximum information method to obtain the overall scores from MIRT domain scores. A simulation study was conducted to investigate the accuracy and effects of using the MIRT matching test response function (TRF) linking procedure to link the five-dimensional domain scores and their overall scores under varying conditions of anchor set lengths, population distribution types, and sample sizes. The results show that the TRF method recovered the domain scores and overall scores well for an anchor set of 10 items, with 2 items in each domain. Domain scores and overall scores have high reliabilities when the correlations between domains are high; reliability is higher than .9 for overall scores and higher than .8 for domain scores under all conditions.
Citation	Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement*, *35*(1), 48–66. https://doi.org/10.1177/0146621610373095
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621610373095

Title	Two Approaches for Using Multiple Anchors in NEAT Equating: A Description and Demonstration
Abstract	Nonequivalent groups with anchor test (NEAT) equating functions that use a single anchor can have accuracy problems when the groups are extremely different and/or when the anchor weakly correlates with the tests being equated. Proposals have been made to address these issues by incorporating more than one anchor into NEAT equating functions. These proposals have not been extensively considered or comparatively evaluated. This study evaluates two proposed approaches for incorporating more than one anchor into NEAT equating functions, poststratification and missing data imputation. The approaches are studied and compared in an example of equating mixed format tests where the use of multiple equating is expected to improve equating. The results show that both approaches produced nearly equivalent equating results but that the poststratification approach has some flexibility and accuracy advantages over imputation in terms of standard errors.
Citation	Moses, T., Deng, W., & Zhang, Y.-L. (2011). Two approaches for using multiple anchors in NEAT equating: A description and demonstration. *Applied Psychological Measurement*, *35*(5), 362–379. https://doi.org/10.1177/0146621611405510
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621611405510

Title	Exploring the Full-Information Bifactor Model in Vertical Scaling with Construct Shift
Abstract	To address the lack of attention to construct shift in item response theory (IRT) vertical scaling, a multigroup, bifactor model was proposed to model the common dimension for all grades and the grade-specific dimensions. Bifactor model estimation accuracy was evaluated through a simulation study with manipulated factors of percentage of common items, sample size, and degree of construct shift. In addition, the unidimensional IRT (UIRT) model, which ignores construct shift, was also estimated to represent current practice. It was found that (a) bifactor models were well recovered overall, though the grade-specific dimensions were not as well recovered as the general dimension; (b) item discrimination parameter estimates were overestimated in UIRT models due to the effect of construct shift; (c) the person parameters of UIRT models were less accurately estimated than those of bifactor models; (d) group mean parameter estimates from UIRT models were less accurate than those of bifactor models; and (e) a large effect due to construct shift was found for the group mean parameter estimates of UIRT models. A real data analysis provided an illustration of how bifactor models can be applied to problems involving vertical scaling with construct shift. General procedures for testing practice were recommended and discussed.
Citation	Li, Y., & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, *36*(1), 3–20. https://doi.org/10.1177/0146621611432864
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621611432864

Title	Effects of Vertical Scaling Methods on Linear Growth Estimation
Abstract	Vertical scaling is necessary to facilitate comparison of scores from test forms of different difficulty levels. It is widely used to enable the tracking of student growth in academic performance over time. Most previous studies on vertical scaling methods assume relatively long tests and large samples. Little is known about their performance when the sample is small or the test is short, challenges that small testing programs often face. This study examined effects of sample size, test length, and choice of item response theory (IRT) models on the performance of IRT-based scaling methods (concurrent calibration, separate calibration with Stocking–Lord, Haebara, Mean/Mean, and Mean/Sigma transformation) in linear growth estimation when the 2-parameter IRT model was appropriate. Results showed that IRT vertical scales could be used for growth estimation without grossly biasing growth parameter estimates when sample size was not large, as long as the test was not too short (≥20 items), although larger sample sizes would generally increase the stability of the growth parameter estimates. The optimal rate of return in total estimation error reduction as a result of increasing sample size appeared to be around 250. Concurrent calibration produced slightly lower total estimation error than separate calibration in the worst combination of short test length (≤20 items) and small sample size ($n \leq 100$), whereas separate calibration, except in the case of the Mean/Sigma method, produced similar or somewhat lower amounts of total error in other conditions.
Citation	Lei, P.-W., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement*, *36*(1), 21–39. https://doi.org/10.1177/0146621611425171
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621611425171

Title	Investigating the Impact of Compromised Anchor Items on IRT Equating Under the Nonequivalent Anchor Test Design
Abstract	The prevalence of high-stakes test scores as a basis for significant decisions necessitates the dissemination of accurate and fair scores. However, the magnitude of these decisions has created an environment in which examinees may be prone to resort to cheating. To reduce the risk of cheating, multiple test forms are commonly administered. When multiple forms are employed, the forms must be equated to account for potential differences in form difficulty. If cheating occurs on one of the forms, the equating procedure may produce inaccurate results. A simulation study was conducted to examine the impact of cheating on item response theory (IRT) true score equating. Recovery of equated scores and scaling constants was assessed for the Stocking–Lord IRT scaling method under various conditions. Results indicated that cheating artificially increased the equated scores of the entire examinee group that was administered the compromised form. Future research should focus on the identification and removal of compromised items.
Citation	Jurich, D. P., DeMars, C. E., & Goodman, J. T. (2012). Investigating the impact of compromised anchor items on IRT equating under the nonequivalent anchor test design. *Applied Psychological Measurement*, *36*(4), 291–308. https://doi.org/10.1177/0146621612445575
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621612445575

Title	Rasch Scale Stability in the Presence of Item Parameter and Trait Drift
Abstract	Testing programs often rely on common-item equating to maintain a single measurement scale across multiple test administrations and multiple years. Changes over time, in the item parameters and the latent trait underlying the scale, can lead to inaccurate score comparisons and misclassifications of examinees. This study examined how instability in a scale and the items composing a scale affects item parameter recovery and classification accuracy. Results showed that a Rasch item response theory scale can maintain near baseline recovery properties if the changes in the latent trait over time are small. The Rasch scale also maintained good recovery of item and person parameters if there was equal item drift in both directions. Under conditions of relatively little item drift and small to moderate periodic changes in the latent trait, a Rasch scale may remain stable for 15 years, ±3. Substantial item drift or large changes in the latent trait can dramatically reduce the longevity of the scale.
Citation	Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*(7), 565–580. https://doi.org/10.1177/0146621612455090
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621612455090

Title	Book Review: Statistical Models for Test Equating, Scaling, and Linking
Abstract	(N/A)
Citation	B, J. G. (2013). Book review: Statistical models for test equating, scaling, and linking. *Applied Psychological Measurement*, *37*(4), 336–339. https://doi.org/10.1177/0146621613476762
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621613476762

Title	Observed Score and True Score Equating Procedures for Multidimensional Item Response Theory
Abstract	The purpose of this research was to develop observed score and true score equating procedures to be used in conjunction with the multidimensional item response theory (MIRT) framework. Three equating procedures—two observed score procedures and one true score procedure—were created and described in detail. One observed score procedure was presented as a direct extension of unidimensional IRT (UIRT) observed score equating and is referred to as the "Full MIRT Observed Score Equating Procedure." The true score procedure and the second observed score procedure incorporated unidimensional approximation procedures to equate exams using UIRT equating principles. These procedures are referred to as the "Unidimensional Approximation of MIRT True Score Equating Procedure" and the "Unidimensional Approximation of MIRT Observed Score Equating Procedure," respectively. Three exams were used to conduct UIRT observed score and true score equating, MIRT observed score and true score equating, and equipercentile equating. The equipercentile equating procedure was conducted for the purpose of comparison because this procedure does not explicitly violate the IRT assumption of unidimensionality. Results indicated that the MIRT equating procedures performed more similarly to the equipercentile equating procedure than the UIRT equating procedures, presumably due to the violation of the unidimensionality assumption under the UIRT equating procedures.
Citation	Brossman, B. G., & Lee, W.-C. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement*, *37*(6), 460–481. https://doi.org/10.1177/0146621613484083
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621613484083

Title	Using a Linear Regression Method to Detect Outliers in IRT Common Item Equating
Abstract	Common test items play an important role in equating alternate test forms under the common item nonequivalent groups design. When the item response theory (IRT) method is applied in equating, inconsistent item parameter estimates among common items can lead to large bias in equated scores. It is prudent to evaluate inconsistency in parameter estimates of common items before conducting IRT equating. The evaluation of inconsistency in parameter estimates is typically achieved through detecting outliers in the common item set. In this study, a linear regression method is proposed as a detection method. The newly proposed method was compared with a traditional method in various conditions. The results of this study confirmed the necessity of detecting and removing outlying common items. The results also show that the newly proposed method performed better than did the traditional method in most conditions.
Citation	He, Y., Cui, Z., Fang, Y., & Chen, H. (2013). Using a linear regression method to detect outliers in IRT common item equating. *Applied Psychological Measurement*, *37*(7), 522–540. https://doi.org/10.1177/0146621613483207
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621613483207

Title	Cross-Validation: An Alternative Bandwidth-Selection Method in Kernel Equating
Abstract	The development of the kernel equating (KE) method enhanced the theory of observed-score equating. In KE, discrete test score distributions are converted into continuous distributions through the use of a Gaussian kernel. Traditionally, the optimal bandwidth for a Gaussian kernel was obtained by minimizing a penalty function. In this article, an alternative bandwidth-selection approach for KE was adopted that uses cross-validation (CV) techniques. The method is illustrated through simulations that were conducted with 188 conditions by varying three factors known to influence equating results; these include sample sizes, score distributions, and methods that involve both equating and bandwidth-selection methods. Four equating procedures were considered: traditional equipercentile equating, which uses linear interpolation to make the test distributions continuous; KE with penalty functions; and KE with two newly proposed CV methods. The results were evaluated based on four criteria: bias in continuizing the distributions (i.e., the difference between the estimated and underlying score distributions), the standard error of equating (SEE), the difference between equated scores, and percent relative error (PRE). Overall, the results demonstrate that KE with the two CV methods outperformed the others—the estimated density functions were less biased and the SEEs and PREs were smaller. Equating differences between the different methods were produced, although they were not large. In addition, the bias issues surrounding kernel methods on sample sizes and the shapes of the distributions were addressed and discussed.
Citation	Liang, T., & von Davier, A. A. (2014). Cross-validation: An alternative bandwidth-selection method in kernel equating. *Applied Psychological Measurement*, *38*(4), 281–295. https://doi.org/10.1177/0146621613518094
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621613518094

Title	Examining Potential Boundary Bias Effects in Kernel Smoothing on Equating: An Introduction for the Adaptive and Epanechnikov Kernels
Abstract	Test equating is a method of making the test scores from different test forms of the same assessment comparable. In the equating process, an important step involves continuizing the discrete score distributions. In traditional observed-score equating, this step is achieved using linear interpolation (or an unscaled uniform kernel). In the kernel equating (KE) process, this continuization process involves Gaussian kernel smoothing. It has been suggested that the choice of bandwidth in kernel smoothing controls the trade-off between variance and bias. In the literature on estimating density functions using kernels, it has also been suggested that the weight of the kernel depends on the sample size, and therefore, the resulting continuous distribution exhibits bias at the endpoints, where the samples are usually smaller. The purpose of this article is (a) to explore the potential effects of atypical scores (spikes) at the extreme ends (high and low) on the KE method in distributions with different degrees of asymmetry using the randomly equivalent groups equating design (Study I), and (b) to introduce the Epanechnikov and adaptive kernels as potential alternative approaches to reducing boundary bias in smoothing (Study II). The beta-binomial model is used to simulate observed scores reflecting a range of different skewed shapes.
Citation	Cid, J. A., & von Davier, A. A. (2015). Examining potential boundary bias effects in kernel smoothing on equating: An introduction for the adaptive and Epanechnikov kernels. *Applied Psychological Measurement*, *39*(3), 208–222. https://doi.org/10.1177/0146621614555901
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621614555901

Title	Kernel Equating Under the Non-Equivalent Groups With Covariates Design
Abstract	When equating two tests, the traditional approach is to use common test takers and/or common items. Here, the idea is to use variables correlated with the test scores (e.g., school grades and other test scores) as a substitute for common items in a non-equivalent groups with covariates (NEC) design. This is performed in the framework of kernel equating and with an extension of the method developed for post-stratification equating in the non-equivalent groups with anchor test design. Real data from a college admissions test were used to illustrate the use of the design. The equated scores from the NEC design were compared with equated scores from the equivalent group (EG) design, that is, equating with no covariates as well as with equated scores when a constructed anchor test was used. The results indicate that the NEC design can produce lower standard errors compared with an EG design. When covariates were used together with an anchor test, the smallest standard errors were obtained over a large range of test scores. The results obtained, that an EG design equating can be improved by adjusting for differences in test score distributions caused by differences in the distribution of covariates, are useful in practice because not all standardized tests have anchor tests.
Citation	Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, *39*(5), 349–361. https://doi.org/10.1177/0146621614567939
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621614567939

Title	Item Response Theory Models for Carry-Over Effect Across Different Scales
Abstract	It is common in educational and psychological tests or social surveys that the same statement is judged on multiple scales. These multiple responses are linked by the same statement, which may cause local dependence. Considering the way a statement is judged on multiple scales, a new class of item response theory (IRT) models is developed to account for the nonrecursive carry-over effect, in which a response can be affected only by its preceding response rather than by a subsequent response. The parameters of the models can be estimated with the freeware WinBUGS. Two simulation studies were conducted to evaluate the parameter recovery of the new models and the consequences of model misspecification. Results showed that the parameters of the new models were recovered fairly well; fitting unnecessarily complicated models to data that did not have the carry-over effect did little harm to parameter estimation; and ignoring the carry-over effect by fitting standard IRT models yielded biased estimates for the item parameters, the correlation between latent traits, and the test reliability. Two empirical examples with parallel design and sequential design are provided to demonstrate the implications and applications of the new models.
Citation	Jin, K.-Y., & Wang, W.-C. (2015). Item response theory models for carry-over effect across different scales. *Applied Psychological Measurement*, *39*(5), 406–425. https://doi.org/10.1177/0146621615572250
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621615572250

Title	New Robust Scale Transformation Methods in the Presence of Outlying Common Items
Abstract	Common items play an important role in item response theory (IRT) true score equating under the common-item nonequivalent groups design. Biased item parameter estimates due to common item outliers can lead to large errors in equated scores. Current methods used to screen for common item outliers mainly focus on the detection and elimination of those items, which may lead to inadequate content representation for the common items. To reduce the impact of inconsistency in item parameter estimates while maintaining content representativeness, the authors propose two robust scale transformation methods based on two weighting methods: the Area-Weighted method and the Least Absolute Values (LAV) method. Results from two simulation studies indicate that these robust scale transformation methods performed as well as the Stocking-Lord method in the absence of common item outliers and, more importantly, outperformed the Stocking-Lord method when a single outlying common item was simulated.
Citation	He, Y., Cui, Z., & Osterlind, S. J. (2015). New robust scale transformation methods in the presence of outlying common items. *Applied Psychological Measurement*, *39*(8), 613–626. https://doi.org/10.1177/0146621615587003
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621615587003

Title	Alternative Linear Item Response Theory Observed-Score Equating Methods
Abstract	Item response theory observed-score equating (IRTOSE) is widely used in many testing programs. The aim of this study was to empirically examine three alternative linear IRTOSE methods compared with the traditional IRTOSE method and to discuss these methods in light of previously suggested alternatives. This contribution is both conceptual, by exploring three alternative methods that fit into the current observed-score equating framework, and empirical by comparing the methods through simulations and with real data. The results show that the local linear (kernel) IRTOSE methods yield low bias and low values on loss measures. However, using only a linear IRTOSE method results in excessive bias and cannot be recommended because of the ease with which IRTOSE with full distributions can be performed. An example using real data showed considerable differences in the equated scores with the alternative methods as well as in comparison with the traditional IRTOSE method. Practical considerations are given in the concluding remarks.
Citation	Wiberg, M. (2016). Alternative linear item response theory observed-score equating methods. *Applied Psychological Measurement*, *40*(3), 180–199. https://doi.org/10.1177/0146621615605089
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621615605089

Title	Bias in Point Estimates and Standard Errors of Mokken's Scalability Coefficients
Abstract	Mokken scale analysis uses three types of scalability coefficients to assess the quality of (a) pairs of items, (b) individual items, and (c) an entire scale. Both the point estimates and the standard errors of the scalability coefficients assume that the sample ordering of the item steps is identical to the population ordering, but due to sampling error, the sample ordering may be incorrect and, consequently, the estimates and the standard errors may be biased. Two simulation studies were used to investigate the bias of the estimates and the standard errors of the scalability coefficients, as well as the coverage of the 95% confidence intervals. Distance between item steps was the most important design factor. In addition, sample size, number of items, number of answer categories, and item discrimination were included in the design. Bias of the standard errors was negligible. Bias of the estimates was largest when all item steps were identical in the population, especially for small sample sizes. Furthermore, bias of the estimates decreased as number of answer categories increased and as item discrimination decreased. Coverage of the 95% confidence intervals was close to .950, but for small sample size coverage deteriorated. Coverage also became poorer as number of items increased, in particular for dichotomous items.
Citation	Kuijpers, R. E., van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2016). Bias in point estimates and standard errors of Mokken's scalability coefficients. *Applied Psychological Measurement*, *40*(5), 331–345. https://doi.org/10.1177/0146621616638500
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621616638500

Title	equate: Observed-Score Linking and Equating in R
Abstract	(N/A)
Citation	Albano, A. D. (2016). equate: Observed-score linking and equating in R. *Applied Psychological Measurement*, *40*(5), 361–362. https://doi.org/10.1177/0146621615620553
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621615620553

Title	Evaluating Anchor-Item Designs for Concurrent Calibration With the GGUM
Abstract	Concurrent calibration using anchor items has proven to be an effective alternative to separate calibration and linking for developing large item banks, which are needed to support continuous testing. In principle, anchor-item designs and estimation methods that have proven effective with dominance item response theory (IRT) models, such as the 3PL model, should also lead to accurate parameter recovery with ideal point IRT models, but surprisingly little research has been devoted to this issue. This study, therefore, had two purposes: (a) to develop software for concurrent calibration with, what is now the most widely used ideal point model, the generalized graded unfolding model (GGUM); (b) to compare the efficacy of different GGUM anchor-item designs and develop empirically based guidelines for practitioners. A Monte Carlo study was conducted to compare the efficacy of three anchor-item designs in vertical and horizontal linking scenarios. The authors found that a block-interlaced design provided the best parameter recovery in nearly all conditions. The implications of these findings for concurrent calibration with the GGUM and practical recommendations for pretest designs involving ideal point computer adaptive testing (CAT) applications are discussed.
Citation	Joo, S.-H., Lee, P., & Stark, S. (2017). Evaluating anchor-item designs for concurrent calibration with the GGUM. *Applied Psychological Measurement*, *41*(2), 83–96. https://doi.org/10.1177/0146621616673997
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621616673997

Title	Linking Methods for the Zinnes–Griggs Pairwise Preference IRT Model
Abstract	Forced-choice item response theory (IRT) models are being more widely used as a way of reducing response biases in noncognitive research and operational testing contexts. As applications have increased, there has been a growing need for methods to link parameters estimated in different examinee groups as a prelude to measurement equivalence testing. This study compared four linking methods for the Zinnes and Griggs (ZG) pairwise preference ideal point model. A Monte Carlo simulation compared test characteristic curve (TCC) linking, item characteristic curve (ICC) linking, mean/mean (M/M) linking, and mean/sigma (M/S) linking. The results indicated that ICC linking and the simpler M/M and M/S methods performed better than TCC linking, and there were no substantial differences among the top three approaches. In addition, in the absence of possible contamination of the common (anchor) item subset due to differential item functioning, five items should be adequate for estimating the metric transformation coefficients. Our article presents the necessary equations for ZG linking and provides recommendations for practitioners who may be interested in developing and using pairwise preference measures for research and selection purposes.
Citation	Lee, P., Joo, S.-H., & Stark, S. (2017). Linking methods for the Zinnes–Griggs pairwise preference IRT model. *Applied Psychological Measurement*, *41*(2), 130–144. https://doi.org/10.1177/0146621616675836
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621616675836

Title	Ability and Prior Distribution Mismatch: An Exploration of Common-Item Linking Methods
Abstract	Linking of two forms is an important task when using item response theory, particularly when two forms are administered to nonequivalent groups. When linking with characteristic curve methods, the ability distribution and weights associated with that distribution can be used to weight observations differently. These are commonly specified as equally spaced intervals from −4 to 4, but other options or distributional forms can be specified. The use of these different distributions and weights of the ability distributions will be explored with a Monte Carlo simulation. Primary simulation conditions will include sample size, number of items, number of common items, ability distribution, and randomly varying population transformation constants. Study results show that the linking weights have little impact on the estimation of the linking constants; however, the underlying ability distribution of examinees does have significant impact. Implications for applied researchers will be discussed.
Citation	LeBeau, B. (2017). Ability and prior distribution mismatch: An exploration of common-item linking methods. *Applied Psychological Measurement*, *41*(7), 545–560. https://doi.org/10.1177/0146621617707508
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621617707508

Title	A Comparative Evaluation of Kernel Equating and Test Characteristic Curve Equating
Abstract	This study compares the kernel equating (KE) and test characteristic curve (TCC) equating methods using the nonequivalent anchor test equating design. In this Monte Carlo study, four independent variables were examined: sample size, test length, average form discrimination, anchor test reliability, and the percentage of anchor items. For each condition, there were 100 replications. To assess the performance of TCC equating and KE, the differences between the examinee parametric true scores and the equated estimated expected true scores were examined. The equated scores were based on the average across replications for each condition. Generally speaking, both KE and TCC equating produced accurate results, although KE tended to perform better than TCC on the parametric true score scale across conditions. Past research and the current study's results seem to indicate that KE should be strongly considered for most equating situations, particularly in light of its flexibility.
Citation	De Ayala, R. J., Smith, B., & Norman Dvorak, R. (2018). A comparative evaluation of kernel equating and test characteristic curve equating. *Applied Psychological Measurement*, *42*(2), 155–168. https://doi.org/10.1177/0146621617712245
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621617712245

Title	Asymptotic Variance of Linking Coefficient Estimators for Polytomous IRT Models
Abstract	In item response theory (IRT), when two groups from different populations take two separate tests, there is a need to link the two ability scales so that the item parameters of the tests are comparable across the groups. To link the two scales, information from common items are utilized to estimate linking coefficients which place the item parameters on the same scale. For polytomous IRT models, the Haebara and Stocking–Lord methods for estimating the linking coefficients have commonly been recommended. However, estimates of the variance for these methods are not available in the literature. In this article, the asymptotic variance of linking coefficients for polytomous IRT models with the Haebara and Stocking–Lord methods are derived. The results are presented in a general form and specific results are given for the generalized partial credit model. Simulations which investigate the accuracy of the derivations under various settings of model complexity and sample size are provided, showing that the derivations are accurate under the conditions considered and that the Haebara and Stocking–Lord methods have superior performance to several moment methods with performance close to that of concurrent calibration.
Citation	Andersson, B. (2018). Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Applied Psychological Measurement*, *42*(3), 192–205. https://doi.org/10.1177/0146621617721249
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621617721249

Title	Book Review: Applying Test Equating Methods: Using R
Abstract	(N/A)
Citation	Battauz, M. (2018). Book review: Applying test equating methods: Using R. *Applied Psychological Measurement*, *42*(7), 590–592. https://doi.org/10.1177/0146621617752992
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621617752992

Title	Linking With External Covariates: Examining Accuracy by Anchor Type, Test Length, Ability Difference, and Sample Size
Abstract	Research has recently demonstrated the use of multiple anchor tests and external covariates to supplement or substitute for common anchor items when linking and equating with nonequivalent groups. This study examines the conditions under which external covariates improve linking and equating accuracy, with internal and external anchor tests of varying lengths and groups of differing abilities. Pseudo forms of a state science test were equated within a resampling study where sample size ranged from 1,000 to 10,000 examinees and anchor tests ranged in length from eight to 20 items, with reading and math scores included as covariates. Frequency estimation linking with an anchor test and external covariate was found to produce the most accurate results under the majority of conditions studied. Practical applications of linking with anchor tests and covariates are discussed.
Citation	Albano, A. D., & Wiberg, M. (2019). Linking with external covariates: Examining accuracy by anchor type, test length, ability difference, and sample size. *Applied Psychological Measurement*, *43*(8), 597–610. https://doi.org/10.1177/0146621618824855
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621618824855

Title	Approximating Bifactor IRT True-Score Equating with a Projective Item Response Model
Abstract	Item response theory (IRT) true-score equating for the bifactor model is often conducted by first numerically integrating out specific factors from the item response function and then applying the unidimensional IRT true-score equating method to the marginalized bifactor model. However, an alternative procedure for obtaining the marginalized bifactor model is through projecting the nuisance dimensions of the bifactor model onto the dominant dimension. Projection, which can be viewed as an approximation to numerical integration, has an advantage over numerical integration in providing item parameters for the marginalized bifactor model; therefore, projection could be used with existing equating software packages that require item parameters. In this paper, IRT true-score equating results obtained with projection are compared to those obtained with numerical integration. Simulation results show that the two procedures provide very similar equating results.
Citation	Kim, K. Y., & Cho, U. H. (2020). Approximating bifactor IRT true-score equating with a projective item response model. *Applied Psychological Measurement*, *44*(3), 215–218. https://doi.org/10.1177/0146621619885903
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621619885903

Title	Evaluating Robust Scale Transformation Methods with Multiple Outlying Common Items Under IRT True Score Equating
Abstract	Item parameter estimates of a common item on a new test form may change abnormally due to reasons such as item overexposure or change of curriculum. A common item, whose change does not fit the pattern implied by the normally behaved common items, is defined as an outlier. Although improving equating accuracy, detecting and eliminating of outliers may cause a content imbalance among common items. Robust scale transformation methods have recently been proposed to solve this problem when only one outlier is present in the data, although it is not uncommon to see multiple outliers in practice. In this simulation study, the authors examined the robust scale transformation methods under conditions where there were multiple outlying common items. Results indicated that the robust scale transformation methods could reduce the influences of multiple outliers on scale transformation and equating. The robust methods performed similarly to a traditional outlier detection and elimination method in terms of reducing the influence of outliers while keeping adequate content balance.
Citation	He, Y., & Cui, Z. (2020). Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Applied Psychological Measurement*, *44*(4), 296–310. https://doi.org/10.1177/0146621619886050
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621619886050

Title	On the Practical Consequences of Misfit in Mokken Scaling
Abstract	Mokken scale analysis is a popular method to evaluate the psychometric quality of clinical and personality questionnaires and their individual items. Although many empirical papers report on the extent to which sets of items form Mokken scales, there is less attention for the effect of violations of commonly used rules of thumb. In this study, the authors investigated the practical consequences of retaining or removing items with psychometric properties that do not comply with these rules of thumb. Using simulated data, they concluded that items with low scalability had some influence on the reliability of test scores, person ordering and selection, and criterion-related validity estimates. Removing the misfitting items from the scale had, in general, a small effect on the outcomes. Although important outcome variables were fairly robust against scale violations in some conditions, authors conclude that researchers should not rely exclusively on algorithms allowing automatic selection of items. In particular, content validity must be taken into account to build sensible psychometric instruments.
Citation	Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2020). On the practical consequences of misfit in Mokken scaling. *Applied Psychological Measurement*, *44*(6), 482–496. https://doi.org/10.1177/0146621620920925
Link	https://journals.sagepub.com/doi/full/10.1177/0146621620920925

Title	Asymptotic Standard Errors of Parameter Scale Transformation Coefficients in Test Equating Under the Nominal Response Model
Abstract	Researchers have developed a characteristic curve procedure to estimate the parameter scale transformation coefficients in test equating under the nominal response model. In the study, the delta method was applied to derive the standard error expressions for computing the standard errors for the estimates of the parameter scale transformation coefficients. This brief report presents the results of a simulation study that examined the accuracy of the derived formulas and compared the performance of this analytical method with that of the multiple imputation method. The results indicated that the standard errors produced by the delta method were very close to the criterion standard errors as well as those yielded by the multiple imputation method under all the simulation conditions.
Citation	Zhang, Z. (2021). Asymptotic standard errors of parameter scale transformation coefficients in test equating under the nominal response model. *Applied Psychological Measurement*, *45*(2), 134–138. https://doi.org/10.1177/0146621620965740
Link	https://journals.sagepub.com/doi/abs/10.1177/0146621620965740

Title	Scale Alignment in the Between-Item Multidimensional Partial Credit Model
Abstract	In between-item multidimensional item response models, it is often desirable to compare individual latent trait estimates across dimensions. These comparisons are only justified if the model dimensions are scaled relative to each other. Traditionally, this scaling is done using approaches such as standardization—fixing the latent mean and standard deviation to 0 and 1 for all dimensions. However, approaches such as standardization do not guarantee that Rasch model properties hold across dimensions. Specifically, for between-item multidimensional Rasch family models, the unique ordering of items holds within dimensions, but not across dimensions. Previously, Feuerstahler and Wilson described the concept of scale alignment, which aims to enforce the unique ordering of items across dimensions by linearly transforming item parameters within dimensions. In this article, we extend the concept of scale alignment to the between-item multidimensional partial credit model and to models fit using incomplete data. We illustrate this method in the context of the Kindergarten Individual Development Survey (KIDS), a multidimensional survey of kindergarten readiness used in the state of Illinois. We also present simulation results that demonstrate the effectiveness of scale alignment in the context of polytomous item response models and missing data.
Citation	Feuerstahler, L., & Wilson, M. (2021). Scale alignment in the between-item multidimensional partial credit model. *Applied Psychological Measurement*, *45*(4), 268–282. https://doi.org/10.1177/01466216211013103
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216211013103

Title	Asymptotic Standard Errors of Generalized Partial Credit Model True Score Equating Using Characteristic Curve Methods
Abstract	In this study, the delta method was applied to estimate the standard errors of the true score equating when using the characteristic curve methods with the generalized partial credit model in test equating under the context of the common-item nonequivalent groups equating design. Simulation studies were further conducted to compare the performance of the delta method with that of the bootstrap method and the multiple imputation method. The results indicated that the standard errors produced by the delta method were very close to the criterion empirical standard errors as well as those yielded by the bootstrap method and the multiple imputation method under all the manipulated conditions.
Citation	Zhang, Z. (2021). Asymptotic standard errors of generalized partial credit model true score equating using characteristic curve methods. *Applied Psychological Measurement*, *45*(5), 331–345. https://doi.org/10.1177/01466216211013101
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216211013101

Title	PROsetta: An R Package for Linking Patient-Reported Outcome Measures
Abstract	A common problem when using a variety of patient-reported outcomes (PROs) for diverse populations and subgroups is establishing a harmonized scale for the incommensurate outcomes. The lack of comparability in metrics (e.g., raw summed scores vs. scaled scores) among different PROs poses practical challenges in studies comparing effects across studies and samples. Linking has long been used for practical benefit in educational testing. Applying various linking techniques to PRO data has a relatively short history; however, in recent years, there has been a surge of published studies on linking PROs and other health outcomes, owing in part to concerted efforts such as the Patient-Reported Outcomes Measurement Information System (PROMIS®) project and the PRO Rosetta Stone (PROsetta Stone®) project (www.prosettastone.org). Many R packages have been developed for linking in educational settings; however, they are not tailored for linking PROs where harmonization of data across clinical studies or settings serves as the main objective. We created the PROsetta package to fill this gap and disseminate a protocol that has been established as a standard practice for linking PROs.
Citation	Choi, S. W., Lim, S., Schalet, B. D., Kaat, A. J., & Cella, D. (2021). PROsetta: An R package for linking patient-reported outcome measures. *Applied Psychological Measurement*, *45*(5), 386–388. https://doi.org/10.1177/01466216211013106
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216211013106

Title	Partial Measurement Invariance: Extending and Evaluating the Cluster Approach for Identifying Anchor Items
Abstract	When measurement invariance does not hold, researchers aim for partial measurement invariance by identifying anchor items that are assumed to be measurement invariant. In this paper, we build on Bechger and Maris's approach for identification of anchor items. Instead of identifying differential item functioning (DIF)-free items, they propose to identify different sets of items that are invariant in item parameters within the same item set. We extend their approach by an additional step in order to allow for identification of homogeneously functioning item sets. We evaluate the performance of the extended cluster approach under various conditions and compare its performance to that of previous approaches, that are the equal-mean difficulty (EMD) approach and the iterative forward approach. We show that the EMD and the iterative forward approaches perform well in conditions with balanced DIF or when DIF is small. In conditions with large and unbalanced DIF, they fail to recover the true group mean differences. With appropriate threshold settings, the cluster approach identified a cluster that resulted in unbiased mean difference estimates in all conditions. Compared to previous approaches, the cluster approach allows for a variety of different assumptions as well as for depicting the uncertainty in the results that stem from the choice of the assumption. Using a real data set, we illustrate how the assumptions of the previous approaches may be incorporated in the cluster approach and how the chosen assumption impacts the results.
Citation	Pohl, S., Schulze, D., & Stets, E. (2021). Partial measurement invariance: Extending and evaluating the cluster approach for identifying anchor items. *Applied Psychological Measurement*, *45*(7–8), 477–493. https://doi.org/10.1177/01466216211042809
Link	https://journals.sagepub.com/doi/full/10.1177/01466216211042809

Title	How Important is the Choice of Bandwidth in Kernel Equating?
Abstract	Kernel equating uses kernel smoothing techniques to continuize the discrete score distributions when equating test scores from an assessment test. The degree of smoothness of the continuous approximations is determined by the bandwidth. Four bandwidth selection methods are currently available for kernel equating, but no thorough comparison has been made between these methods. The overall aim is to compare these four methods together with two additional methods based on cross-validation in a simulation study. Both equivalent and non-equivalent group designs are used and the number of test takers, test length, and score distributions are all varied. The results show that sample size and test length are important factors for equating accuracy and precision. However, all bandwidth selection methods perform similarly with regards to the mean squared error and the differences in terms of equated scores are small, suggesting that the choice of bandwidth is not critical. The different bandwidth selection methods are also illustrated using real testing data from a college admissions test. Practical implications of the results from the simulation study and the empirical study are discussed.
Citation	Wallin, G., Häggström, J., & Wiberg, M. (2021). How important is the choice of bandwidth in kernel equating? *Applied Psychological Measurement*, *45*(7–8), 518–535. https://doi.org/10.1177/01466216211040486
Link	https://journals.sagepub.com/doi/full/10.1177/01466216211040486

Title	Scale Linking for the Testlet Item Response Theory Model
Abstract	In their 2005 paper, Li and her colleagues proposed a test response function (TRF) linking method for a two-parameter testlet model and used a genetic algorithm to find minimization solutions for the linking coefficients. In the present paper the linking task for a three-parameter testlet model is formulated from the perspective of bi-factor modeling, and three linking methods for the model are presented: the TRF, mean/least squares (MLS), and item response function (IRF) methods. Simulations are conducted to compare the TRF method using a genetic algorithm with the TRF and IRF methods using a quasi-Newton algorithm and the MLS method. The results indicate that the IRF, MLS, and TRF methods perform very well, well, and poorly, respectively, in estimating the linking coefficients associated with testlet effects, that the use of genetic algorithms offers little improvement to the TRF method, and that the minimization function for the TRF method is not as well-structured as that for the IRF method.
Citation	Kim, S., & Kolen, M. J. (2022). Scale linking for the testlet item response theory model. *Applied Psychological Measurement*, *46*(2), 79–97. https://doi.org/10.1177/01466216211063234
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216211063234

Title	Standard Errors of Kernel Equating: Accounting for Bandwidth Estimation
Abstract	In standardized testing, equating is used to ensure comparability of test scores across multiple test administrations. One equipercentile observed-score equating method is kernel equating, where an essential step is to obtain continuous approximations to the discrete score distributions by applying a kernel with a smoothing bandwidth parameter. When estimating the bandwidth, additional variability is introduced which is currently not accounted for when calculating the standard errors of equating. This poses a threat to the accuracy of the standard errors of equating. In this study, the asymptotic variance of the bandwidth parameter estimator is derived and a modified method for calculating the standard error of equating that accounts for the bandwidth estimation variability is introduced for the equivalent groups design. A simulation study is used to verify the derivations and confirm the accuracy of the modified method across several sample sizes and test lengths as compared to the existing method and the Monte Carlo standard error of equating estimates. The results show that the modified standard errors of equating are accurate under the considered conditions. Furthermore, the modified and the existing methods produce similar results which suggest that the bandwidth variability impact on the standard error of equating is minimal.
Citation	Marcq, K., & Andersson, B. (2022). Standard errors of kernel equating: Accounting for bandwidth estimation. *Applied Psychological Measurement*, *46*(3), 200–218. https://doi.org/10.1177/01466216211066601
Link	https://journals.sagepub.com/doi/full/10.1177/01466216211066601

Title	BayMDS: An R Package for Bayesian Multidimensional Scaling and Choice of Dimension
Abstract	(N/A)
Citation	Oh, M.-S., & Lee, E.-K. (2022). BayMDS: An R package for Bayesian multidimensional scaling and choice of dimension. *Applied Psychological Measurement*, *46*(3), 250–251. https://doi.org/10.1177/01466216221084219
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216221084219

Title	Item Response Theory True Score Equating for the Bifactor Model Under the Common-Item Nonequivalent Groups Design
Abstract	Applying item response theory (IRT) true score equating to multidimensional IRT models is not straightforward due to the one-to-many relationship between a true score and latent variables. Under the common-item nonequivalent groups design, the purpose of the current study was to introduce two IRT true score equating procedures that adopted different dimension reduction strategies for the bifactor model. The first procedure, which was referred to as the integration procedure, linked the latent variable scales for the bifactor model and integrated out the specific factors from the item response function of the bifactor model. Then, IRT true score equating was applied to the marginalized bifactor model. The second procedure, which was referred to as the PIRT-based procedure, projected the specific dimensions onto the general dimension to obtain a locally dependent unidimensional IRT (UIRT) model and linked the scales of the UIRT model, followed by the application of IRT true score equating to the locally dependent UIRT model. Equating results obtained with the two equating procedures along with those obtained with the unidimensional three-parameter logistic (3PL) model were compared using both simulated and real data. In general, the integration and PIRT-based procedures provided equating results that were not practically different. Furthermore, the equating results produced by the two bifactor-based procedures became more accurate than the results returned by the 3PL model as tests became more multidimensional.
Citation	Kim, K. Y. (2022). Item response theory true score equating for the bifactor model under the common-item nonequivalent groups design. *Applied Psychological Measurement*, *46*(6), 479–493. https://doi.org/10.1177/01466216221108995
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216221108995

Title	Characterizing Sampling Variability for Item Response Theory Scale Scores in a Fixed-Parameter Calibrated Projection Design
Abstract	A common practice of linking uses estimated item parameters to calculate projected scores. This procedure fails to account for the carry-over sampling variability. Neglecting sampling variability could consequently lead to understated uncertainty for Item Response Theory (IRT) scale scores. To address the issue, we apply a Multiple Imputation (MI) approach to adjust the Posterior Standard Deviations of IRT scale scores. The MI procedure involves drawing multiple sets of plausible values from an approximate sampling distribution of the estimated item parameters. When two scales to be linked were previously calibrated, item parameters can be fixed at their original published scales, and the latent variable means and covariances of the two scales can then be estimated conditional on the fixed item parameters. The conditional estimation procedure is a special case of Restricted Recalibration (RR), in which the asymptotic sampling distribution of estimated parameters follows from the general theory of pseudo Maximum Likelihood (ML) estimation. We evaluate the combination of RR and MI by a simulation study to examine the impact of carry-over sampling variability under various simulation conditions. We also illustrate how to apply the proposed method to real data by revisiting Thissen et al. (2015).
Citation	Xu, S., & Liu, Y. (2022). Characterizing sampling variability for item response theory scale scores in a fixed-parameter calibrated projection design. *Applied Psychological Measurement*, *46*(6), 509–528. https://doi.org/10.1177/01466216221108136
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216221108136

Title	Application of Sampling Variance of Item Response Theory Parameter Estimates in Detecting Outliers in Common Item Equating
Abstract	In common item equating, the existence of item outliers may impact the accuracy of equating results and bring significant ramifications to the validity of test score interpretations. Therefore, common item equating should involve a screening process to flag outlying items and exclude them from the common item set before equating is conducted. The current simulation study demonstrated that the sampling variance associated with the item response theory (IRT) item parameter estimates can help detect outliers in the common items under the 2-PL and 3-PL IRT models. The results showed the proposed sampling variance statistic (SV) outperformed the traditional displacement method with cutoff values of 0.3 and 0.5 along a variety of evaluation criteria. Based on the favorable results, item outlier detection statistics based on estimated sampling variability warrant further consideration in both research and practice.
Citation	Liu, C., & Jurich, D. (2022). Application of sampling variance of item response theory parameter estimates in detecting outliers in common item equating. *Applied Psychological Measurement*, *46*(6), 529–547. https://doi.org/10.1177/01466216221108122
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216221108122

Title	Investigating the Effect of Differential Rapid Guessing on Population Invariance in Equating
Abstract	Score equating is an essential tool in improving the fairness of test score interpretations when employing multiple test forms. To ensure that the equating functions used to connect scores from one form to another are valid, they must be invariant across different populations of examinees. Given that equating is used in many low-stakes testing programs, examinees' test-taking effort should be considered carefully when evaluating population invariance in equating, particularly as the occurrence of rapid guessing (RG) has been found to differ across subgroups. To this end, the current study investigated whether differential RG rates between subgroups can lead to incorrect inferences concerning population invariance in test equating. A simulation was built to generate data for two examinee subgroups (one more motivated than the other) administered two alternative forms of multiple-choice items. The rate of RG and ability characteristics of rapid guessers were manipulated. Results showed that as RG responses increased, false positive and false negative inferences of equating invariance were respectively observed at the lower and upper ends of the observed score scale. This result was supported by an empirical analysis of an international assessment. These findings suggest that RG should be investigated and documented prior to test equating, especially in low-stakes assessment contexts. A failure to do so may lead to incorrect inferences concerning fairness in equating.
Citation	Deng, J., & Rios, J. A. (2022). Investigating the effect of differential rapid guessing on population invariance in equating. *Applied Psychological Measurement*, *46*(7), 589–604. https://doi.org/10.1177/01466216221108991
Link	https://journals.sagepub.com/doi/abs/10.1177/01466216221108991

Educational Measurement: Issues and Practice (EM:IP): 2010–2022

Title	Measurement, Sampling, and Equating Errors in Large-Scale Assessments
Abstract	In large-scale assessments, such as state-wide testing programs, national sample-based assessments, and international comparative studies, there are many steps involved in the measurement and reporting of student achievement. There are always sources of inaccuracies in each of the steps. It is of interest to identify the source and magnitude of the errors in the measurement process that may threaten the validity of the final results. Assessment designers can then improve the assessment quality by focusing on areas that pose the highest threats to the results. This paper discusses the relative magnitudes of three main sources of error with reference to the objectives of assessment programs: measurement error, sampling error, and equating error. A number of examples from large-scale assessments are used to illustrate these errors and their impact on the results. The paper concludes by making a number of recommendations that could lead to an improvement of the accuracies of large-scale assessment results.
Citation	Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. https://doi.org/10.1111/j.1745-3992.2010.00190.x
Link	https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1745-3992.2010.00190.x

Title	Scaling: An ITEMS Module
Abstract	Scaling is the process of constructing a score scale that associates numbers or other ordered indicators with the performance of examinees. Scaling typically is conducted to aid users in interpreting test results. This module describes different types of raw scores and scale scores, illustrates how to incorporate various sources of information into a score scale, and introduces vertical scaling and its related designs and methodologies as a special type of scaling. After completion of this module, the reader should be able to understand the relationship between various types of raw scores, understand the relationship between raw scores and scale scores, construct a scale with desired properties, evaluate an existing score scale, understand how content and standards information are built into a scale, and understand how vertical scales are developed and used in practice.
Citation	Tong, Y., & Kolen, M. J. (2010). Scaling: an ITEMS module. *Educational Measurement: Issues and Practice*, *29*(4), 39–48. https://doi.org/10.1111/j.1745-3992.2010.00192.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2010.00192.x

Title	Equating Subscores under the Nonequivalent Anchor Test (NEAT) Design
Abstract	The study examined two approaches for equating subscores. They are (1) equating subscores using internal common items as the anchor to conduct the equating, and (2) equating subscores using equated and scaled total scores as the anchor to conduct the equating. Since equated total scores are comparable across the new and old forms, they can be used as an anchor to equate the subscores. Both chained linear and chained equipercentile methods were used. Data from two tests were used to conduct the study and results showed that when more internal common items were available (i.e., 10–12 items), then using common items to equate the subscores is preferable. However, when the number of common items is very small (i.e., five to six items), then using total scaled scores to equate the subscores is preferable. For both tests, not equating (i.e., using raw subscores) is not reasonable as it resulted in a considerable amount of bias.
Citation	Puhan, G., & Liang, L. (2011). Equating subscores under the nonequivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice*, *30*(1), 23–35. https://doi.org/10.1111/j.1745-3992.2010.00197.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2010.00197.x

Title	Psychometric Properties of Raw and Scale Scores on Mixed-Format Tests
Abstract	This paper illustrates that the psychometric properties of scores and scales that are used with mixed-format educational tests can impact the use and interpretation of the scores that are reported to examinees. Psychometric properties that include reliability and conditional standard errors of measurement are considered in this paper. The focus is on mixed-format tests in situations for which raw scores are integer-weighted sums of item scores. Four associated real-data examples include (a) effects of weights associated with each item type on reliability, (b) comparison of psychometric properties of different scale scores, (c) evaluation of the equity property of equating, and (d) comparison of the use of unidimensional and multidimensional procedures for evaluating psychometric properties. Throughout the paper, and especially in the conclusion section, the examples are related to issues associated with test interpretation and test use.
Citation	Kolen, M. J., & Lee, W.-C. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice*, *30*(2), 15–24. https://doi.org/10.1111/j.1745-3992.2011.00201.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2011.00201.x

Title	An NCME Instructional Module on Population Invariance in Linking and Equating
Abstract	A goal for any linking or equating of two or more tests is that the linking function be invariant to the population used in conducting the linking or equating. Violations of population invariance in linking and equating jeopardize the fairness and validity of test scores, and pose particular problems for test-based accountability programs that require schools, districts, and states to report annual progress on academic indicators disaggregated by demographic group membership. This instructional module provides a comprehensive overview of population invariance in linking and equating and the relevant methodology developed for evaluating violations of invariance. A numeric example is used to illustrate the comparative properties of available methods, and important considerations for evaluating population invariance in linking and equating are presented.
Citation	Huggins, A. C., & Penfield, R. D. (2012). An NCME instructional module on population invariance in linking and equating. *Educational Measurement: Issues and Practice*, *31*(1), 27–40. https://doi.org/10.1111/j.1745-3992.2011.00225.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2011.00225.x

Title	The Philosophical Aspects of IRT Equating: Modeling Drift to Evaluate Cohort Growth in Large-Scale Assessments
Abstract	Calibration and equating is the quintessential necessity for most large-scale educational assessments. However, there are instances when no consideration is given to the equating process in terms of context and substantive realization, and the methods used in its execution. In the view of the authors, equating is not merely an exhibit of the statistical methodology, but it is also a reflection of the thought process undertaken in its execution. For example, there is hardly any discussion in literature of the ideological differences in the selection of an equating method. Furthermore, there is little evidence of modeling cohort growth through an identification and use of construct-relevant linking items' drift, using the common item nonequivalent group equating design. In this article, the authors philosophically justify the use of Huynh's statistical method for the identification of construct-relevant outliers in the linking pool. The article also dispels the perception of scale instability associated with the inclusion of construct-relevant outliers in the linking item pool and concludes that an appreciation of the rationale used in the selection of the equating method, together with the use of linking items in modeling cohort growth, can be beneficial to the practitioners.
Citation	Taherbhai, H., & Seo, D. (2013). The philosophical aspects of IRT equating: Modeling drift to evaluate cohort growth in large-scale assessments. *Educational Measurement: Issues and Practice*, *32*(1), 2–14. https://doi.org/10.1111/emip.12000
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12000

Title	Quantifying Error and Uncertainty Reductions in Scaling Functions: An ITEMS Module
Abstract	This module describes and extends X-to-Y regression measures that have been proposed for use in the assessment of X-to-Y scaling and equating results. Measures are developed that are similar to those based on prediction error in regression analyses but that are directly suited to interests in scaling and equating evaluations. The regression and scaling function measures are compared in terms of their uncertainty reductions, error variances, and the contribution of true score and measurement error variances to the total error variances. The measures are also demonstrated as applied to an assessment of scaling results for a math test and a reading test. The results of these analyses illustrate the similarity of the regression and scaling measures for scaling situations when the tests have a correlation of at least .80, and also show the extent to which the measures can be adequate summaries of nonlinear regression and nonlinear scaling functions, and of heteroskedastic errors. After reading this module, readers will have a comprehensive understanding of the purposes, uses, and differences of regression and scaling functions.
Citation	Moses, T. (2014). Quantifying error and uncertainty reductions in scaling functions: An ITEMS module. *Educational Measurement: Issues and Practice*, *33*(2), 29–40. https://doi.org/10.1111/emip.12032
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12032

Title	On the Choice of Anchor Tests in Equating
Abstract	The choice of anchor tests is crucial in applications of the nonequivalent groups with anchor test design of equating. Sinharay and Holland (2006, 2007) suggested "miditests," which are anchor tests that are content-representative and have the same mean item difficulty as the total test but have a smaller spread of item difficulties. Sinharay and Holland (2006, 2007), Cho, Wall, Lee, and Harris (2010), Fitzpatrick and Skorupski (2016), Liu, Sinharay, Holland, Curley, and Feigenbaum (2011a), Liu, Sinharay, Holland, Feigenbaum, and Curley (2011b), and Yi (2009) found the miditests to lead to better equating than minitests, which are representative of the total test with respect to content and difficulty. However, these findings recently came into question as Trierweiler, Lewis, and Smith (2016) concluded, based on a comparison of correlation coefficients of miditests and minitests with the total test, that making an anchor test a miditest does not generally increase the anchor to total score correlation and recommended the continuation of the practice of using minitests over miditests. Their recommendation raises the question, "Should miditests continue to be considered in practice?" This note defends the miditests by citing literature that favors miditests and then by showing that miditests perform as well as the minitests in most realistic situations considered in Trierweiler et al. (2016), which implies that miditests should continue to be seriously considered by equating practitioners.
Citation	Sinharay, S. (2018). On the choice of anchor tests in equating. *Educational Measurement: Issues and Practice*, *37*(2), 64–69. https://doi.org/10.1111/emip.12175
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12175

Title	How Robust Are Cross-Country Comparisons of PISA Scores to the Scaling Model Used?
Abstract	The Programme for International Student Assessment (PISA) is an important international study of 15-olds' knowledge and skills. New results are released every 3 years, and have a substantial impact upon education policy. Yet, despite its influence, the methodology underpinning PISA has received significant criticism. Much of this criticism has focused upon the psychometric scaling model used to create the proficiency scores. The aim of this article is to therefore investigate the robustness of cross-country comparisons of PISA scores to subtle changes to the underlying scaling model used. This includes the specification of the item-response model, whether the difficulty and discrimination of items are allowed to vary across countries (item-by-country interactions) and how test questions not reached by pupils are treated. Our key finding is that these technical choices make little substantive difference to the overall country-level results.
Citation	Jerrim, J., Parker, P., Choi, A., Chmielewski, A. K., Sälzer, C., & Shure, N. (2018). How robust are cross-country comparisons of PISA scores to the scaling model used?. *Educational Measurement: Issues and Practice*, *37*(4), 28–39. https://doi.org/10.1111/emip.12211
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12211

Title	A Comparison of Two Alternate Scaling Approaches Employed for Task Analyses in Credentialing Examination Development
Abstract	Credentialing examination developers rely on task (job) analyses for establishing inventories of task and knowledge areas in which competency is required for safe and successful practice in target occupations. There are many ways in which task-related information may be gathered from practitioner ratings, each with its own advantage and limitation. Two of the myriad alternative task analysis rating approaches are compared in situ: one establishing relative task saliency through a single scale of rated importance and another employing a composite of several independent scales. Outcomes regarding tasks ranked by two practitioner groups are compared. A relatively high degree of association is observed between tasks ranked through each approach, yielding comparable, though not identical examination blueprints.
Citation	Fidler, J. R., & Risk, N. M. (2019). A comparison of two alternate scaling approaches employed for task analyses in credentialing examination development. *Educational Measurement: Issues and Practice*, *38*(1), 78–86. https://doi.org/10.1111/emip.12200
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12200

Title	Evaluating Population Invariance of Test Equating During the COVID-19 Pandemic
Abstract	Population invariance is a desirable property of test equating which might not hold when significant changes occur in the test population, such as those brought about by the COVID-19 pandemic. This research aims to investigate whether equating functions are reasonably invariant when the test population is impacted by the pandemic. Based on pseudo-test forms constructed from an operational form administered in the springs of 2019, 2020, and 2021, this study conducted preequating and postequating using different data collection designs and different sample sizes based on each year's data. Raw-to-scale score conversion tables from each equating and the group means after applying these conversions were compared with those from a criterion equating, that is, single group design postequating based on the 2019 data. Within each design and sample size condition, the magnitude of the differences between the 2021 equating and the criterion equating was mostly similar to the magnitude of differences between the 2020 equating and the criterion, indicating a reasonable extent of invariance in equating results. Nevertheless, some equating designs showed slightly less invariance than others.
Citation	Li, D., & Kapoor, S. (2022). Evaluating population invariance of test equating during the COVID-19 pandemic. *Educational Measurement: Issues and Practice*, *41*(1), 33–41. https://doi.org/10.1111/emip.12489
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12489

Title	Adjusting for Ability Differences of Equating Samples When Randomization Is Suboptimal
Abstract	Test equating requires collecting data to link the scores from different forms of a test. Problems arise when equating samples are not equivalent and the test forms to be linked share no common items by which to measure or adjust for the group nonequivalence. Using data from five operational test forms, we created five pairs of research forms for each form, such that the equating relationship between each pair was known. Then we compared five approaches to adjusting for group nonequivalence in a situation where not only was group equivalence questionable, but the number of common items was small. We used a resampling approach to evaluate the linking accuracy of group adjustment using sample weights via minimum discriminant information adjustment (MDIA) using test takers' collateral (demographic) information, a weak anchor of only three items, or a mix of both. Overall, the use of both sample weights via MDIA and a weak anchor produced the most accurate result, while the direct (random groups) linking method assuming group equivalence produced the least accurate result due to nontrivial bias. For all five research forms, using both collateral information and anchor items only marginally improved linking accuracy compared to using the weak anchor alone.
Citation	Kim, S., & Walker, M. E. (2022). Adjusting for ability differences of equating samples when randomization is suboptimal. *Educational Measurement: Issues and Practice*, *41*(3), 26–37. https://doi.org/10.1111/emip.12506
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12506

Title	Digital Module 29: Multidimensional Item Response Theory Equating
Abstract	In this digital ITEMS module, Dr. Stella Kim provides an overview of multidimensional item response theory (MIRT) equating. Traditional unidimensional item response theory (IRT) equating methods impose the sometimes untenable restriction on data that only a single ability is assessed. This module discusses potential sources of multidimensionality and presents potential consequences of multidimensionality on equating. To remedy these effects, MIRT equating can be used as a viable alternative to traditional methods of IRT equating. In conducting MIRT equating, the choice of an appropriate MIRT model is necessary, and thus the module describes several existing MIRT models and illustrates each using hypothetical examples. After a brief description of MIRT models, an extensive review of the current literature is presented to identify gaps in the literature on MIRT equating. Then, the steps for conducting MIRT observed-score equating are described. Finally, the module discusses practical considerations in applying MIRT equating to testing practices and suggests potential areas of research for future studies.
Citation	Kim, S. Y. (2022). Digital module 29: Multidimensional item response theory equating. *Educational Measurement: Issues and Practice*, *41*(3), 85–86. https://doi.org/10.1111/emip.12525
Link	https://onlinelibrary.wiley.com/doi/10.1111/emip.12525

Educational and Psychological Measurement (EPM): 2010–2022

Title	Assessing Fit and Dimensionality in Least Squares Metric Multidimensional Scaling Using Akaike's Information Criterion
Abstract	Akaike's information criterion is suggested as a tool for evaluating fit and dimensionality in metric multidimensional scaling that uses least squares methods of estimation. This criterion combines the least squares loss function with the number of estimated parameters. Numerical examples are presented. The results from analyses of both simulation data and real data demonstrate the utility of the Akaike's information criterion in identifying the best approximating models in multidimensional scaling analyses.
Citation	Ding, C. S., & Davison, M. L. (2010). Assessing fit and dimensionality in least squares metric multidimensional scaling using Akaike's Information Criterion. *Educational and Psychological Measurement*, *70*(2), 199–214. https://doi.org/10.1177/0013164409344554
Link	https:// https://journals.sagepub.com/doi/10.1177/0013164409344554

Title	A Brief Report on How Impossible Scores Affect Smoothing and Equating
Abstract	Equating under the external anchor design is frequently conducted using scaled scores on the anchor test. However, scaled scores often lead to the unique problem of creating zero frequencies in the score distribution because there may not always be a one-to-one correspondence between raw and scaled scores. For example, raw scores of 17 and 18 may correspond to scaled scores of 150 and 153, thereby creating zero frequencies for scaled scores of 151 and 152. These gaps in the frequency distribution may adversely affect smoothing and equating. This study examines the effect of these zero frequencies on log-linear smoothing of score distributions and final equating results. Results suggest that although smoothing is significantly affected by the presence of these zero frequencies, the impact on the actual equating results is minimal.
Citation	Puhan, G., von Davier, A. A., & Gupta, S. (2010). A brief report on how impossible scores affect smoothing and equating. *Educational and Psychological Measurement*, *70*(6), 953–960. https://doi.org/10.1177/0013164410382731
Link	https://journals.sagepub.com/doi/10.1177/0013164410382731

Title	Observed Score Equating Using a Mini-Version Anchor and an Anchor with Less Spread of Difficulty: A Comparison Study
Abstract	Two different types of anchors are investigated in this study: a mini-version anchor and an anchor that has a less spread of difficulty than the tests to be equated. The latter is referred to as a midi anchor. The impact of these two different types of anchors on observed score equating are evaluated and compared with respect to systematic error (bias), random equating error (SEE), and total equating error (RMSE) using SAT operational data. The results suggest that the overall bias, SEE, and RMSE when the midi anchor is used are either smaller than or very similar to those when the mini anchor test is used. The findings suggest that a midi anchor test would be preferred to a mini anchor test if equating accuracy at the ends of the score scale is not a primary concern.
Citation	Liu, J., Sinharay, S., Holland, P., Feigenbaum, M., & Curley, E. (2011). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement*, *71*(2), 346–361. https://doi.org/10.1177/0013164410375571
Link	https://journals.sagepub.com/doi/abs/10.1177/0013164410375571

Title	The Long-Term Sustainability of Different Item Response Theory Scaling Methods
Abstract	This article investigates the accuracy of examinee classification into performance categories and the estimation of the theta parameter for several item response theory (IRT) scaling techniques when applied to six administrations of a test. Previous research has investigated only two administrations; however, many testing programs equate tests across multiple administrations. As such, this article seeks to examine the long-term sustainability of IRT scaling methods. Three different types of shifts in the ability distribution were examined: no change, a mean shift, and a change in skewness. Haebara, Stocking and Lord, mean—sigma, mean—mean, and fixed common item parameter (FCIP) scaling were compared relative to bias, root mean square error, and classification of examinees into performance categories. Results indicate that FCIP may be the most suitable for complex changes in examinee performance, whereas the methods performed quite similarly for simple changes.
Citation	Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educational and Psychological Measurement*, *71*(2), 362–379. https://doi.org/10.1177/0013164410375111
Link	https://journals.sagepub.com/doi/abs/10.1177/0013164410375111

Title	Evaluating Ranking Strategies in Assessing Change When the Measures Differ Across Time
Abstract	In this study, a ranking strategy was evaluated for comparing subgroups' change using identical, equated, and nonidentical measures. Four empirical data sets were evaluated, each of which contained examinees' scores on two occasions, where the two occasions' scores were obtained on a single identical measure, on two equated tests, and on two nonidentical measures. The two subgroups' rates of change were compared based on ranked nonidentical measures, on raw and ranked equated measures, and on a raw and ranked identical measure. The results of comparing subgroups' change were similar when based on the nonidentical measures and on the identical and equated measures. Additional evaluations using simulated data demonstrated that the ranking strategy proposed for nonidentical measures is accurate, especially when the subgroups were large, the difference between the subgroups' change was large, and scores obtained on the measure(s) were highly correlated across occasions. The statistical power of the proposed ranking method is slightly reduced because of the tendency of the nonidentical measures to have relatively low correlations.
Citation	Moses, T., & Kim, S. (2012). Evaluating ranking strategies in assessing change when the Measures differ across time. *Educational and Psychological Measurement*, *72*(1), 78–98. https://doi.org/10.1177/0013164411405651
Link	https://journals.sagepub.com/doi/full/10.1177/0013164411405651

Title	Nominal Weights Mean Equating: A Method for Very Small Samples
Abstract	The authors introduced nominal weights mean equating, a simplified version of Tucker equating, as an alternative for dealing with very small samples. The authors then conducted three simulation studies to compare nominal weights mean equating to six other equating methods under the nonequivalent groups anchor test design with sample sizes of 20, 50, and 80 examinees. Results showed that nominal weights mean equating was generally the most effective. Nominal weights mean equating was, furthermore, never among the least effective methods in any condition, indicating its utility across a wide variety of contexts. Circle-arc equating, another recently developed method, also showed a great deal of promise. The identity function (i.e., no equating) was adequate only when test forms were nearly equivalent in difficulty.
Citation	Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement*, *72*(4), 608–628. https://doi.org/10.1177/0013164411428609
Link	https://journals.sagepub.com/doi/full/10.1177/0013164411428609

Title	An Application of Explanatory Item Response Modeling for Model-Based Proficiency Scaling
Abstract	The article compares three different methods to estimate effects of task characteristics and to use these estimates for model-based proficiency scaling: prediction of item difficulties from the Rasch model, the linear logistic test model (LLTM), and an LLTM including random item effects (LLTM+e). The methods are applied to empirical data from a German large-scale study of reading comprehension in English as a foreign language (N = 10,543). A priori defined task characteristics were used as predictors for item difficulty; the estimated effects were used to define thresholds between proficiency levels. The comparison of results indicates that the LLTM is too restrictive; the Rasch model and the LLTM+e yield similar results in terms of implications for scale anchoring.
Citation	Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*(4), 665–686. https://doi.org/10.1177/0013164411430707
Link	https://journals.sagepub.com/doi/full/10.1177/0013164411430707

Title	Linking Cut-Scores Given Changes in the Decision-Making Process, Administration Time, and Proportions of Item Types Between Successive Administrations of a Test for a Large-Scale Assessment Program
Abstract	There is a continuing tension in testing programs to equate forms and maintain score scales and at the same time allow for changing conditions in the educational system, such as curriculum shifts or practical limits on testing time. When such changes occur, psychometric staff members are challenged to develop linking methods that allow for comparable reporting but meet requirements for psychometric rigor. This article describes a method addressing such shifts in testing programs. The application of the method is demonstrated on a large-scale educational testing program that had changes in test length, content distribution, and decision-making process. The method used to accomplish the linkage was to develop a pseudo test from the items included in the longer test before the change that was designed to mimic the test after the change. The linking of the tests using the pseudo test process resulted in a percentage of successful students that was similar to the percentages obtained prior to the changes. The linked scores were treated as comparable rather than equated scores.
Citation	Radwan, N., Reckase, M. D., & Rogers, W. T. (2013). Linking cut-scores given changes in the decision-making process, administration time, and proportions of item types between successive administrations of a test for a large-scale assessment program. *Educational and Psychological Measurement*, *73*(1), 125–142. https://doi.org/10.1177/0013164412448652
Link	https://journals.sagepub.com/doi/full/10.1177/0013164412448652

Title	A New Method for Analyzing Content Validity Data Using Multidimensional Scaling
Abstract	Validity evidence based on test content is of essential importance in educational testing. One source for such evidence is an alignment study, which helps evaluate the congruence between tested objectives and those specified in the curriculum. However, the results of an alignment study do not always sufficiently capture the degree to which a test adequately represents the intended content domain. In this study, we present and evaluate a method for analyzing alignment data that uses multidimensional scaling to determine whether the dimensions underlying subject matter experts' (SME) ratings of test items conform to the dimensions delineated in the test specifications. The results suggest this procedure provides an enhanced view of content domain representation by portraying the content similarities among items along continuous, rather than discrete, dimensions. The additional types of information provided by this procedure, in relation to that provided by traditional analysis of SME data, are discussed.
Citation	Li, X., & Sireci, S. G. (2013). A new method for analyzing content validity data using multidimensional scaling. *Educational and Psychological Measurement*, *73*(3), 365–385. https://doi.org/10.1177/0013164412473825
Link	https://journals.sagepub.com/doi/full/10.1177/0013164412473825

Title	Effects of Item Parameter Drift on Vertical Scaling With the Nonequivalent Groups With Anchor Test (NEAT) Design
Abstract	The authors explored the effects of drifting common items on vertical scaling within the higher order framework of item parameter drift (IPD). The results showed that if IPD occurred between a pair of test levels, the scaling performance started to deviate from the ideal state, as indicated by bias of scaling. When there were two items drifting with 0.5 logits, IPD could have a substantial effect on vertical scaling. Although IPD had little impact on the recoveries of the parameters on the whole developmental scale, its effects on the recoveries of some parameters by separate grade and grade pair were manifest. Specifically, the mean achievement estimates became worse conditional on the pair of test levels between which IPD occurred, and the estimations of grade-to-grade growth and effect size were distorted for the grade pair corresponding to the test pair that involved IPD. Neither the estimation of standard deviation of the achievement nor the grade-to-grade variability was influenced by IPD.
Citation	Ye, M., & Xin, T. (2014). Effects of item parameter drift on vertical scaling with the nonequivalent groups with anchor test (NEAT) design. *Educational and Psychological Measurement*, *74*(2), 227–235. https://doi.org/10.1177/0013164413513024
Link	https://journals.sagepub.com/doi/full/10.1177/0013164413513024

Title	Alternative Smoothing and Scaling Strategies for Weighted Composite Scores
Abstract	In this study, smoothing and scaling approaches are compared for estimating subscore-to-composite scaling results involving composites computed as rounded and weighted combinations of subscores. The considered smoothing and scaling approaches included those based on raw data, on smoothing the bivariate distribution of the subscores, on smoothing the bivariate distribution of the subscore and weighted composite, and two weighted averages of the raw and smoothed marginal distributions. Results from simulations showed that the approaches differed in terms of their estimation accuracy for scaling situations with smaller and larger sample sizes, and on weighted composite distributions of varied complexity.
Citation	Moses, T. (2014). Alternative smoothing and scaling strategies for weighted composite scores. *Educational and Psychological Measurement*, *74*(3), 516–536. https://doi.org/10.1177/0013164413507725
Link	https://journals.sagepub.com/doi/full/10.1177/0013164413507725

Title	The Effect of Differential Item Functioning in Anchor Items on Population Invariance of Equating
Abstract	Invariant relationships in the internal mechanisms of estimating achievement scores on educational tests serve as the basis for concluding that a particular test is fair with respect to statistical bias concerns. Equating invariance and differential item functioning are both concerned with invariant relationships yet are treated separately in the psychometric literature. Connecting these two facets of statistical invariance is critical for developing a holistic definition of fairness in educational measurement, for fostering a deeper understanding of the nature and causes of equating invariance and a lack thereof, and for providing practitioners with guidelines for addressing reported score-level equity concerns. This study hypothesizes that differential item functioning manifested in anchor items of an assessment will have an effect on equating dependence. Findings show that when anchor item differential item functioning varies across forms in a differential manner across subpopulations, population invariance of equating can be compromised.
Citation	Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement*, *74*(4), 627–658. https://doi.org/10.1177/0013164413506222
Link	https://journals.sagepub.com/doi/full/10.1177/0013164413506222

Title	Enhancing a Short Measure of Big Five Personality Traits With Bayesian Scaling
Abstract	A study in a university clinic/laboratory investigated adaptive Bayesian scaling as a supplement to interpretation of scores on the Mini-IPIP. A "probability of belonging" in categories of low, medium, or high on each of the Big Five traits was calculated after each item response and continued until all items had been used or until a predetermined criteria for the posterior probability has been obtained. The study found higher levels of correspondence with the IPIP-50 score categories using the adaptive Bayesian scaling than with the Mini-IPIP alone. The number of additional items ranged from a mean of 2.9 to 12.5 contingent on the level of certainty desired.
Citation	Jones, W. P. (2014). Enhancing a short measure of big five personality traits with Bayesian scaling. *Educational and Psychological Measurement*, *74*(6), 1049–1066. https://doi.org/10.1177/0013164414525040
Link	https://journals.sagepub.com/doi/full/10.1177/0013164414525040

Title	Controlling Guessing Bias in the Dichotomous Rasch Model Applied to a Large-Scale, Vertically Scaled Testing Program
Abstract	Recent research has shown how the statistical bias in Rasch model difficulty estimates induced by guessing in multiple-choice items can be eliminated. Using vertical scaling of a high-profile national reading test, it is shown that the dominant effect of removing such bias is a nonlinear change in the unit of scale across the continuum. The consequence is that the proficiencies of the more proficient students are increased relative to those of the less proficient. Not controlling the guessing bias underestimates the progress of students across 7 years of schooling with important educational implications.
Citation	Andrich, D., Marais, I., & Humphry, S. M. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educational and Psychological Measurement*, *76*(3), 412–435. https://doi.org/10.1177/0013164415594202
Link	https://journals.sagepub.com/doi/full/10.1177/0013164415594202

Title	An Approach to Scoring and Equating Tests With Binary Items: Piloting With Large-Scale Assessments
Abstract	This article describes an approach to test scoring, referred to as delta scoring (D-scoring), for tests with dichotomously scored items. The D-scoring uses information from item response theory (IRT) calibration to facilitate computations and interpretations in the context of large-scale assessments. The D-score is computed from the examinee's response vector, which is weighted by the expected difficulties (not "easiness") of the test items. The expected difficulty of each item is obtained as an analytic function of its IRT parameters. The D-scores are independent of the sample of test-takers as they are based on expected item difficulties. It is shown that the D-scale performs a good bit better than the IRT logit scale by criteria of scale intervalness. To equate D-scales, it is sufficient to rescale the item parameters, thus avoiding tedious and error-prone procedures of mapping test characteristic curves under the method of IRT true score equating, which is often used in the practice of large-scale testing. The proposed D-scaling proved promising under its current piloting with large-scale assessments and the hope is that it can efficiently complement IRT procedures in the practice of large-scale testing in the field of education and psychology.
Citation	Dimitrov, D. M. (2016). An approach to scoring and equating tests with binary items: Piloting with large-scale assessments. *Educational and Psychological Measurement*, *76*(6), 954–975. https://doi.org/10.1177/0013164416631100
Link	https://journals.sagepub.com/doi/full/10.1177/0013164416631100

Title	An examination of Alternative Multidimensional Scaling Techniques
Abstract	The purpose of this study is to compare alternative multidimensional scaling (MDS) methods for constraining the stimuli on the circumference of a circle and on the surface of a sphere. Specifically, the existing MDS-T method for plotting the stimuli on the circumference of a circle is applied, and its extension is proposed for constraining the stimuli on the surface of a sphere. The data analyzed come from previous research and concerns Maslach and Jackson's burnout syndrome and Holland's vocational personality types. The configurations for the same data on the circle and the sphere shared similarities but also had differences, that is, the general item-groupings were the same but most of the differences across the two methods resulted in more meaningful interpretations for the three-dimensional configuration. Furthermore, in most cases, items and/or scales could be better discriminated from each other on the sphere.
Citation	Papazoglou, S., & Mylonas, K. (2017). An examination of alternative multidimensional scaling techniques. *Educational and Psychological Measurement*, *77*(3), 429–448. https://doi.org/10.1177/0013164416661823
Link	https://journals.sagepub.com/doi/full/10.1177/0013164416661823

Title	Simultaneous Linking of Cross-Informant and Longitudinal Data Involving Positive Family Relationships
Abstract	Measurement invariance is a prerequisite when comparing different groups of individuals or when studying a group of individuals across time. This assures that the same construct is assessed without measurement artifacts. This investigation applied a novel approach of simultaneous parameter linking to cross-sectional and longitudinal measures of the construct of positive family relationships. Previously, a scale to measure this construct in mothers was developed longitudinally using the nominal response model of item response theory. In this study, this methodology was conducted for the first time to develop such a scale for children. The data for both informants derived from the Fullerton Longitudinal Study and encompassed 9 annual assessments spanning 8-years (age 9-17 years). This permitted linking across informants studied concurrently and prospectively. This procedure minimized measurement error, furnished a common metric across informants and time and established measurement invariance. Resulting thetas revealed a significant degree of concordance between informants across assessment waves as well as stability of individual differences for both informants over time. This psychometric investigation is unique because it simultaneously established invariance of a construct across informants and time. Implications for future research are discussed.
Citation	Preston, K. S. J., Gottfried, A. W., Park, J. J., Manapat, P. D., Gottfried, A. E., & Oliver, P. H. (2018). Simultaneous linking of cross-informant and longitudinal data involving positive family relationships. *Educational and Psychological Measurement*, *78*(3), 409–429. https://doi.org/10.1177/0013164417690198
Link	https://journals.sagepub.com/doi/full/10.1177/0013164417690198

Title	The Stabilizing Influences of Linking Set Size and Model–Data Fit in Sparse Rater-Mediated Assessment Networks
Abstract	Previous research includes frequent admonitions regarding the importance of establishing connectivity in data collection designs prior to the application of Rasch models. However, details regarding the influence of characteristics of the linking sets used to establish connections among facets, such as locations on the latent variable, model–data fit, and sample size, have not been thoroughly explored. These considerations are particularly important in assessment systems that involve large proportions of missing data (i.e., sparse designs) and are associated with high-stakes decisions, such as teacher evaluations based on teaching observations. The purpose of this study is to explore the influence of characteristics of linking sets in sparsely connected rating designs on examinee, rater, and task estimates. A simulation design whose characteristics were intended to reflect practical large-scale assessment networks with sparse connections were used to consider the influence of locations on the latent variable, model–data fit, and sample size within linking sets on the stability and model–data fit of estimates. Results suggested that parameter estimates for examinee and task facets are quite robust to modifications in the size, model–data fit, and latent-variable location of the link. Parameter estimates for the rater, while still quite robust, are more sensitive to reductions in link size. The implications are discussed as they relate to research, theory, and practice.
Citation	Wind, S. A., & Jones, E. (2018). The stabilizing influences of linking set size and model–data fit in sparse rater-mediated assessment networks. *Educational and Psychological Measurement*, *78*(4), 679–707. https://doi.org/10.1177/0013164417703733
Link	https://journals.sagepub.com/doi/full/10.1177/0013164417703733

Title	The Delta-Scoring Method of Tests With Binary Items: A Note on True Score Estimation and Equating
Abstract	This article presents some new developments in the methodology of an approach to scoring and equating of tests with binary items, referred to as delta scoring (D-scoring), which is under piloting with large-scale assessments at the National Center for Assessment in Saudi Arabia. This presentation builds on a previous work on delta scoring and adds procedures for scaling and equating, item response function, and estimation of true values and standard errors of D scores. Also, unlike the previous work on this topic, where D-scoring involves estimates of item and person parameters in the framework of item response theory, the approach presented here does not require item response theory calibration.
Citation	Dimitrov, D. M. (2018). The delta-scoring method of tests with binary items: A note on true score estimation and equating. *Educational and Psychological Measurement*, *78*(5), 805–825. https://doi.org/10.1177/0013164417724187
Link	https://journals.sagepub.com/doi/full/10.1177/0013164417724187

Title	Evaluating the Accuracy of the Empirical Item Characteristic Curve Preequating Method in the Presence of Test Speededness
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.
Citation	Qiu, Y., & Huggins-Manley, A. C. (2019). Evaluating the accuracy of the empirical item characteristic curve preequating method in the presence of test speededness. *Educational and Psychological Measurement*, *79*(2), 288–309. https://doi.org/10.1177/0013164418777854
Link	https://journals.sagepub.com/doi/full/10.1177/0013164418777854

Title	Simple-Structure Multidimensional Item Response Theory Equating for Multidimensional Tests
Abstract	A theoretical and conceptual framework for true-score equating using a simple-structure multidimensional item response theory (SS-MIRT) model is developed. A true-score equating method, referred to as the SS-MIRT true-score equating (SMT) procedure, also is developed. SS-MIRT has several advantages over other complex multidimensional item response theory models including improved efficiency in estimation and straightforward interpretability. The performance of the SMT procedure was examined and evaluated through four studies using different data types. In these studies, results from the SMT procedure were compared with results from four other equating methods to assess the relative benefits of SMT compared with the other procedures. In general, SMT showed more accurate equating results compared with the traditional unidimensional IRT (UIRT) equating when the data were multidimensional. More accurate performance of SMT over UIRT true-score equating was consistently observed across the studies, which supports the benefits of a multidimensional approach in equating for multidimensional data. Also, SMT performed similarly to a SS-MIRT observed score method across all studies.
Citation	Kim, S. Y., Lee, W.-C., & Kolen, M. J. (2020). Simple-structure multidimensional item response theory equating for multidimensional tests. *Educational and Psychological Measurement*, *80*(1), 91–125. https://doi.org/10.1177/0013164419854208
Link	https://journals.sagepub.com/doi/full/10.1177/0013164419854208

Title	A Method for Examining the Equating of Psychometric Scales and Tests: An Application Using Dementia Screening Test Batteries
Abstract	Equating of psychometric scales and tests is frequently required and conducted in educational, behavioral, and clinical research. Construct comparability or equivalence between measuring instruments is a necessary condition for making decisions about linking and equating resulting scores. This article is concerned with a widely applicable method for examining if two scales or tests cannot be equated. A latent variable modeling method is discussed that can be used to evaluate whether the tests or parts thereof measure latent constructs that are distinct from each other. The approach can be routinely used before an equating procedure is undertaken, in order to assess whether equating could be meaningfully carried out to begin with. The procedure is readily applicable in empirical research using popular software. The method is illustrated with data from dementia screening test batteries administered as part of two studies designed to evaluate a wide range of biomarkers throughout the process of normal aging to dementia or Alzheimer's disease.
Citation	Dowling, N. M., Raykov, T., & Marcoulides, G. A. (2020). A method for examining the equating of psychometric scales and tests: An application using dementia screening test batteries. *Educational and Psychological Measurement*, *80*(1), 199–209. https://doi.org/10.1177/0013164418775785
Link	https://journals.sagepub.com/doi/full/10.1177/0013164418775785

Title	Rasch Versus Classical Equating in the Context of Small Sample Sizes
Abstract	Equating and scaling in the context of small sample exams, such as credentialing exams for highly specialized professions, has received increased attention in recent research. Investigators have proposed a variety of both classical and Rasch-based approaches to the problem. This study attempts to extend past research by (1) directly comparing classical and Rasch techniques of equating exam scores when sample sizes are small ($N \leq 100$ per exam form) and (2) attempting to pool multiple forms' worth of data to improve estimation in the Rasch framework. We simulated multiple years of a small-sample exam program by resampling from a larger certification exam program's real data. Results showed that combining multiple administrations' worth of data via the Rasch model can lead to more accurate equating compared to classical methods designed to work well in small samples. WINSTEPS-based Rasch methods that used multiple exam forms' data worked better than Bayesian Markov Chain Monte Carlo methods, as the prior distribution used to estimate the item difficulty parameters biased predicted scores when there were difficulty differences between exam forms.
Citation	Babcock, B., & Hodge, K. J. (2020). Rasch versus classical equating in the context of small sample sizes. *Educational and Psychological Measurement*, *80*(3), 499–521. https://doi.org/10.1177/0013164419878483
Link	https://journals.sagepub.com/doi/full/10.1177/0013164419878483

Journal of Educational and Behavioral Statistics (JEBS): 2010–2022

Title	Standard Errors of Equating Differences: Prior Developments, Extensions, and Simulations
Abstract	The purpose of this article was to extend the use of standard errors for equated score differences (SEEDs) to traditional equating functions. The SEEDs are described in terms of their original proposal for kernel equating functions and extended so that SEEDs for traditional linear and traditional equipercentile equating functions can be computed. These developments provide new understandings of the relationships between kernel and traditional equating functions that expand on prior developments of SEEDs and of standard errors of equating functions. The developments are demonstrated for an equivalent groups equating situation. The accuracies of the SEEDs are evaluated in simulations conducted using an equivalent groups equating example.
Citation	Moses, T., & Zhang, W. (2011). Standard errors of equating differences: Prior developments, extensions, and simulations. *Journal of Educational and Behavioral Statistics*, *36*(6), 779–803. https://doi.org/10.3102/1076998610396892
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998610396892

Title	The Analysis of Likert Scales Using State Multipoles: An Application of Quantum Methods to Behavioral Sciences Data
Abstract	Though ubiquitous, Likert scaling's traditional mode of analysis is often unable to uncover all of the valid information in a data set. Here, the authors discuss a solution to this problem based on methodology developed by quantum physicists: the state multipole method. The authors demonstrate the relative ease and value of this method by examining college students' endorsement of one possible cause of prejudice: segregation. Though the mean level of students' endorsement did not differ among ethnic groups, an examination of state multipoles showed that African Americans had a level of polarization in their endorsement that was not reflected by Hispanics or European Americans. This result could not have been obtained with the traditional approach and demonstrates the new method's utility for social science research.
Citation	Camparo, J., & Camparo, L. B. (2013). The analysis of Likert scales using state multipoles: An application of quantum methods to behavioral sciences data. *Journal of Educational and Behavioral Statistics*, *38*(1), 81–101. https://doi.org/10.3102/1076998611431084
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998611431084

Title	The Gains From Vertical Scaling
Abstract	It is often assumed that a vertical scale is necessary when value-added models depend upon the gain scores of students across two or more points in time. This article examines the conditions under which the scale transformations associated with the vertical scaling process would be expected to have a significant impact on normative interpretations using gain scores. It is shown that this will depend upon the extent to which adopting a particular vertical scaling approach leads to a large degree of scale shrinkage (decreases in score variability over time). Empirical data are used to compare school-level gain scores computed as a function of different vertical scales transformed to represent increasing, decreasing, and constant trends in score variability across grades. A pragmatic approach is also presented to assess the departure of a given vertical scale from a scale with ideal equal-interval properties. Finally, longitudinal data are used to illustrate a case when the availability of a vertical scale will be most important: when questions are being posed about the magnitudes of student-level growth trajectories.
Citation	Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, *38*(6), 551–576. https://doi.org/10.3102/1076998613508317
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998613508317

Title	Equating Without an Anchor for Nonequivalent Groups of Examinees
Abstract	An equating procedure for a testing program with evolving distribution of examinee profiles is developed. No anchor is available because the original scoring scheme was based on expert judgment of the item difficulties. Pairs of examinees from two administrations are formed by matching on coarsened propensity scores derived from a set of background variables. These two subsets of scores are then equated, treating the associated sets of test performances as equivalent. The method is applied to the scores in 2 years of a testing program for admission to tertiary education.
Citation	Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics*, *40*(3), 227–253. https://doi.org/10.3102/1076998615574773
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998615574773

Title	Pseudo-Equivalent Groups and Linking
Abstract	Adjustment by minimum discriminant information provides an approach to linking test forms in the case of a nonequivalent groups design with no satisfactory common items. This approach employs background information on individual examinees in each administration so that weighted samples of examinees form pseudo-equivalent groups in the sense that they resemble samples from equivalent groups. Linking methods for equivalent groups are then applied to the weighted samples. To illustrate the approach, 29 administrations from a testing program are linked via the method of pseudo-equivalent groups. Because the forms used are currently linked by use of kernel equating, it is possible to compare the reasonableness of results from pseudo-equivalent groups to results from kernel equating.
Citation	Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273. https://doi.org/10.3102/1076998615574772
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998615574772

Title	The New (Educational) Statistics: Properties of Scales That Matter
Abstract	(N/A)
Citation	Ho, A. D. (2016). The new (educational) statistics: Properties of scales that matter. *Journal of Educational and Behavioral Statistics*, *41*(1), 94–99. https://doi.org/10.3102/1076998615621302
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998615621302

Title	Response Styles in Rating Scales: Simultaneous Modeling of Content-Related Effects and the Tendency to Middle or Extreme Categories
Abstract	Heterogeneity in response styles can affect the conclusions drawn from rating scale data. In particular, biased estimates can be expected if one ignores a tendency to middle categories or to extreme categories. An adjacent categories model is proposed that simultaneously models the content-related effects and the heterogeneity in response styles. By accounting for response styles, it provides a simple remedy for the bias that occurs if the response style is ignored. The model allows to include explanatory variables that have a content-related effect as well as an effect on the response style. A visualization tool is developed that makes the interpretation of effects easily accessible. The proposed model is embedded into the framework of multivariate generalized linear model, which entails that common estimation and inference tools can be used. Existing software can be used to fit the model, which makes it easy to apply.
Citation	Tutz, G., & Berger, M. (2016). Response styles in rating scales: Simultaneous modeling of content-related effects and the tendency to middle or extreme categories. *Journal of Educational and Behavioral Statistics*, *41*(3), 239–268. https://doi.org/10.3102/1076998616636850
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998616636850

Title	Estimating Linking Functions for Response Model Parameters
Abstract	Parameter linking in item response theory is generally necessary to adjust for differences between the true values for the same item and ability parameters due to the use of different identifiability restrictions in different calibrations. The research reported in this article explores a precision-weighted (PW) approach to the problem of estimating the linking functions for the common dichotomous logistic response models. Asymptotic standard errors (ASEs) of linking for the new approach are derived and compared to those of the mean/mean and mean/sigma linking methods to which it has a superficial similarity and to the Haebara and Stocking and Lord response function methods. Empirical examples from a few recent linking studies are presented. It is demonstrated that the new approach has smaller ASE than the mean/mean and mean/sigma methods and comparable ASE to the response function methods. However, when some of the item parameters have large estimation error relative to the others, all current methods appear to violate the rather obvious requirement of monotone decrease in ASE with the number of common items in the linking design while the ASE of the PW method demonstrates monotone decrease with the number of common items. The PW method also has the benefits of simple calculation and an ASE which is additive in the contribution of each item, useful for optimal linking design. We conclude that the proposed approach to estimating linking parameters holds promise and warrants further research.
Citation	Barrett, M. D., & van der Linden, W. J. (2019). Estimating linking functions for response model parameters. *Journal of Educational and Behavioral Statistics*, *44*(2), 180–209. https://doi.org/10.3102/1076998618808576
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998618808576

Title	Kernel Equating Using Propensity Scores for Nonequivalent Groups
Abstract	When equating two test forms, the equated scores will be biased if the test groups differ in ability. To adjust for the ability imbalance between nonequivalent groups, a set of common items is often used. When no common items are available, it has been suggested to use covariates correlated with the test scores instead. In this article, we reduce the covariates to a propensity score and equate the test forms with respect to this score. The propensity score is incorporated within the kernel equating framework using poststratification and chained equating. The methods are evaluated using real college admissions test data and through a simulation study. The results show that propensity scores give an increased equating precision in comparison with the equivalent groups design and a smaller mean squared error than by using the covariates directly. Practical implications are also discussed.
Citation	Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics*, *44*(4), 390–414. https://doi.org/10.3102/1076998619838226
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998619838226

Title	Lord's Equity Theorem Revisited
Abstract	Lord's (1980) equity theorem claims observed-score equating to be possible only when two test forms are perfectly reliable or strictly parallel. An analysis of its proof reveals use of an incorrect statistical assumption. The assumption does not invalidate the theorem itself though, which can be shown to follow directly from the discrete nature of the equating problem it addresses. But, surprisingly, an obvious relaxation of the problem is enough to obtain exactly the opposite result: As long as two test forms measure the same ability, they can always be equated, no matter their reliability, degree of parallelness, or even difference in length. Also, in spite of its lack of validity, the original proof of Lord's theorem has an important interim result directly applicable to the problem of assembling a new test form pre-equated to an old form.
Citation	van der Linden, W. J. (2019). Lord's equity theorem revisited. *Journal of Educational and Behavioral Statistics*, *44*(4), 415–430. https://doi.org/10.3102/1076998619837627
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998619837627

Title	A Scaled Threshold Model for Measuring Extreme Response Style
Abstract	Extreme response style is the tendency of individuals to prefer the extreme categories of a rating scale irrespective of item content. It has been shown repeatedly that individual response style differences affect the reliability and validity of item responses and should, therefore, be considered carefully. To account for extreme response style (ERS) in ordered categorical item responses, it has been proposed to model responder-specific sets of category thresholds in connection with established polytomous item response models. An elegant approach to achieve this is to introduce a responder-specific scaling factor that modifies intervals between thresholds. By individually expanding or contracting intervals between thresholds, preferences for selecting either the outer or inner response categories can be modeled. However, for a responder-specific scaling factor to appropriately account for ERS, there are two important aspects that have not been considered previously and which, if ignored, will lead to questionable model properties. Specifically, the centering of threshold parameters and the type of category probability logit need to be considered carefully. In the present article, a scaled threshold model is proposed, which accounts for these considerations. Instructions on model fitting are given together with SAS PROC NLMIXED program code, and the model's application and interpretation is demonstrated using simulation studies and two empirical examples.
Citation	Lubbe, D., & Schuster, C. (2020). A scaled threshold model for measuring extreme response style. *Journal of Educational and Behavioral Statistics*, *45*(1), 86–107. https://doi.org/10.3102/1076998619859541
Link	https://journals.sagepub.com/doi/full/10.3102/1076998619859541

Title	A Bayesian Nonparametric Latent Approach for Score Distributions in Test Equating
Abstract	Equating is a family of statistical models and methods used to adjust scores on different test forms so that they can be comparable and used interchangeably. Equated scores are obtained estimating the equating transformation function, which maps the scores on the scale of one test form into their equivalents on the scale of other one. All the statistical models that have been proposed for estimating this function are based on continuous approximations of the score distributions, leading to equated scores lying on a continuous scale even though score scales are usually subsets of the integer numbers (e.g., the total number of correct answers). In this article, we develop a new equating method from which equated scores defined on the original discrete scale are obtained. Considering scores as ordinal random variables, we propose a continuous latent variable formulation to perform an equipercentile-like equating based on a Bayesian nonparametric model for score distributions. The proposed model is applied to simulated and real data collected under an equivalent group design. Some methods to assess the performance of our model are also discussed. Compared with discrete versions of equated scored obtained from traditional equating methods, the results show that the proposed method has better performance.
Citation	Varas, I. M., González, J., & Quintana, F. A. (2020). A Bayesian nonparametric latent approach for score distributions in test equating. *Journal of Educational and Behavioral Statistics*, *45*(6), 639–666. https://doi.org/10.3102/1076998620907381
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998620907381

Title	Validation Methods for Aggregate-Level Test Scale Linking: A Case Study Mapping School District Test Score Distributions to a Common Scale
Abstract	Linking score scales across different tests is considered speculative and fraught, even at the aggregate level. We introduce and illustrate validation methods for aggregate linkages, using the challenge of linking U.S. school district average test scores across states as a motivating example. We show that aggregate linkages can be validated both directly and indirectly under certain conditions such as when the scores for at least some target units (districts) are available on a common test (e.g., the National Assessment of Educational Progress). We introduce precision-adjusted random effects models to estimate linking error, for populations and for subpopulations, for averages and for progress over time. These models allow us to distinguish linking error from sampling variability and illustrate how linking error plays a larger role in aggregates with smaller sample sizes. Assuming that target districts generalize to the full population of districts, we can show that standard errors for district means are generally less than .2 standard deviation units, leading to reliabilities above .7 for roughly 90% of districts. We also show how sources of imprecision and linking error contribute to both within- and between-state district comparisons within versus between states. This approach is applicable whenever the essential counterfactual question—"what would means/variance/progress for the aggregate units be, had students taken the other test?"—can be answered directly for at least some of the units.
Citation	Reardon, S. F., Kalogrides, D., & Ho, A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics*, *46*(2), 138–167. https://doi.org/10.3102/1076998619874089
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998619874089

Title	Commentary on Reardon, Kalogrides, and Ho's "Validation Methods for Aggregate-Level Test Scale Linking: A Case Study Mapping School District Test Score Distributions to a Common Scale"
Abstract	The studies presented by Reardon, Kalogrides, and Ho provide preliminary support for a National Assessment of Educational Progress–based aggregate linking of state assessments when used for research purposes. In this commentary, I suggest future efforts to explore possible sources of district-level bias, evaluation of predictive accuracy at the state level, and a better understanding of the performance of the linking when applied to the inevitable nonrepresentative district samples that will be encountered in research studies.
Citation	Bolt, D. (2021). Commentary on Reardon, Kalogrides, and Ho's "Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale." *Journal of Educational and Behavioral Statistics*, *46*(2), 168–172. https://doi.org/10.3102/1076998620948267
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998620948267

Title	Commentary on "Validation Methods for Aggregate-Level Test Scale Linking: A Case Study Mapping School District Test Score Distributions to a Common Scale"
Abstract	This paper begins by setting the linking methods of Reardon, Kalogrides, and Ho in the broader literature on linking. Trends in the validity data suggest that there may be a conditional bias in the estimates of district means, but the data in the article are not conclusive on this point. Further, the data used in their case study might support the validity of the methods only over a limited range of the ability continuum. Applications of the method are then discussed. Contrary to the title, the application of the linking results is not limited to aggregate-level data. Because the potential application is so broad, further research is needed on issues such as the possibility of conditional bias and the validity of estimates over the full range of possible values. Validity is not a dichotomous concept where validity exists or it does not. The evidence reported by Reardon et al. provides substantial, but incomplete, support for the validity of the linked measures in this case study.
Citation	Davison, M. L. (2021). Commentary on "Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale." *Journal of Educational and Behavioral Statistics*, *46*(2), 173–186. https://doi.org/10.3102/1076998620949172
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998620949172

Title	Aggregate-Level Test-Scale Linking: A New Solution for an Old Problem?
Abstract	The Reardon, Kalogrides, and Ho article on validation methods for aggregate-level test scale linking is an attempt to validate a district-level scale aligning procedure that appears to be a new solution to an old problem. Their aligning procedure uses the National Assessment of Educational Progress (NAEP) scale to piece together a patchwork of data structures from different tests of different constructs obtained under different administration conditions and used in different ways by different states. In this article, we critique their linking and validation efforts. Our critique has three components. First, we review the recommendations for linking state assessments to NAEP from several studies and commentaries to provide background from which to interpret Reardon et al.'s validation attempts. Second, we provide a replication of the Reardon et al. empirical validations of its proposed linking procedure to demonstrate that correlations between district means on two test scores can be high even when (1) the constructs being measured by the tests are different and (2) the district-level means estimated using the Reardon et al. linking approach can differ substantially from actual district-level means. Then, we suggest additional checks for construct similarity and subpopulation invariance from other concordance studies that could be used to assess whether the inferences made by Reardon et al. are warranted. Finally, until such checks are made, we urge cautious use of the results of the Reardon et al. results.
Citation	Moses, T., & Dorans, N. J. (2021). Aggregate-level test-scale linking: A new solution for an old problem? *Journal of Educational and Behavioral Statistics*, *46*(2), 187–202. https://doi.org/10.3102/1076998620960089
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998620960089

Title	Validation Methods for Aggregate-Level Test Scale Linking: A Rejoinder
Abstract	(N/A)
Citation	Ho, A. D., Reardon, S. F., & Kalogrides, D. (2021). Validation methods for aggregate-level test scale linking: A rejoinder. *Journal of Educational and Behavioral Statistics*, *46*(2), 209–218. https://doi.org/10.3102/1076998621994540
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998621994540

Title	Commentary on “Validation Methods for Aggregate-Level Test Scale Linking: A Case Study Mapping School District Test Score Distributions to a Common Scale”
Abstract	In this commentary, I share my perspective on the goals of assessments in general, on linking assessments that were developed according to different specifications and for different purposes, and I propose several considerations for the authors and the readers. This brief commentary is structured around three perspectives (1) the context of this research, (2) the methodology proposed here, and (3) the consequences for applied research.
Citation	von Davier, A. A. (2021). Commentary on “validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale.” *Journal of Educational and Behavioral Statistics*, *46*(2), 203–208. https://doi.org/10.3102/1076998620956668
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998620956668

Title	A Rating Scale Mixture Model to Account for the Tendency to Middle and Extreme Categories
Abstract	A mixture of logit models is proposed that discriminates between responses to rating questions that are affected by a tendency to prefer middle or extremes of the scale regardless of the content of the item (response styles) and purely content-driven preferences. Explanatory variables are used to characterize the content-driven way of answering as well as the tendency to middle or extreme categories. The proposed model is extended to account for the presence of response styles in the case of several items, and the association among responses is described, both when they are content driven or dictated by response styles. In addition, stochastic orderings, related to the tendency to select middle or extreme categories, are introduced and investigated. A simulation study describes the effectiveness of the proposed model, and an application to a questionnaire on attitudes toward ethnic minorities illustrates the applicability of the modeling approach.
Citation	Colombi, R., Giordano, S., & Tutz, G. (2021). A rating scale mixture model to account for the tendency to middle and extreme categories. *Journal of Educational and Behavioral Statistics*, *46*(6), 682–716. https://doi.org/10.3102/1076998621992554
Link	https://journals.sagepub.com/doi/abs/10.3102/1076998621992554

Title	Mean Comparisons of Many Groups in the Presence of DIF: An Evaluation of Linking and Concurrent Scaling Approaches
Abstract	One of the primary goals of international large-scale assessments in education is the comparison of country means in student achievement. This article introduces a framework for discussing differential item functioning (DIF) for such mean comparisons. We compare three different linking methods: concurrent scaling based on full invariance, concurrent scaling based on partial invariance using the RMSD statistic, and robust and nonrobust linking approaches based on separate scaling. Furthermore, we analytically derive the bias in the country means of different linking methods in the presence of DIF. In a simulation study, we show that the partial invariance and robust linking approaches provide less biased country means than the full invariance approach in the case of biased items.
Citation	Robitzsch, A., & Lüdtke, O. (2022). Mean comparisons of many groups in the presence of DIF: an evaluation of linking and concurrent scaling approaches. *Journal of Educational and Behavioral Statistics*, *47*(1), 36–68. https://doi.org/10.3102/10769986211017479
Link	https://journals.sagepub.com/doi/full/10.3102/10769986211017479

Title	What Is Actually Equated in “Test Equating”? A Didactic Note
Abstract	The current literature on test equating generally defines it as the process necessary to obtain score comparability between different test forms. The definition is in contrast with Lord’s foundational paper which viewed equating as the process required to obtain comparability of measurement scale between forms. The distinction between the notions of scale and score is not trivial. The difference is explained by connecting these notions with standard statistical concepts as probability experiment, sample space, and random variable. The probability experiment underlying equating test forms with random scores immediately gives us the equating transformation as a function mapping the scale of one form into the other and thus supports the point of view taken by Lord. However, both Lord’s view and the current literature appear to rely on the idea of an experiment with random examinees which implies a different notion of test scores. It is shown how an explicit choice between the two experiments is not just important for our theoretical understanding of key notions in test equating but also has important practical consequences.
Citation	der Linden, W. J. van. (2022). What is actually equated in “test equating”? A didactic note. *Journal of Educational and Behavioral Statistics*, *47*(3), 353–362. https://doi.org/10.3102/10769986211072308
Link	https://journals.sagepub.com/doi/full/10.3102/10769986211072308

Title	A Critical View on the NEAT Equating Design: Statistical Modeling and Identifiability Problems
Abstract	The nonequivalent groups with anchor test (NEAT) design is widely used in test equating. Under this design, two groups of examinees are administered different test forms with each test form containing a subset of common items. Because test takers from different groups are assigned only one test form, missing score data emerge by design rendering some of the score distributions unavailable. The partially observed score data formally lead to an identifiability problem, which has not been recognized as such in the equating literature and has been considered from different perspectives, all of them making different assumptions in order to estimate the unidentified score distributions. In this article, we formally specify the statistical model underlying the NEAT design and unveil the lack of identifiability of the parameters of interest that compose the equating transformation. We use the theory of partial identification to show alternatives to traditional practices that have been proposed to identify the score distributions when conducting equating under the NEAT design.
Citation	San Martín, E., & González, J. (2022). A critical view on the NEAT equating design: statistical modeling and identifiability problems. *Journal of Educational and Behavioral Statistics*, *47*(4), 406–437. https://doi.org/10.3102/10769986221090609
Link	https://journals.sagepub.com/doi/abs/10.3102/10769986221090609

Journal of Educational Measurement (JEM): 2010–2022

Title	Comparisons among Designs for Equating Mixed-Format Tests in Large-Scale Assessments
Abstract	In this study we examined variations of the nonequivalent groups equating design for tests containing both multiple-choice (MC) and constructed-response (CR) items to determine which design was most effective in producing equivalent scores across the two tests to be equated. Using data from a large-scale exam, this study investigated the use of anchor CR item rescoring (known as trend scoring) in the context of classical equating methods. Four linking designs were examined: an anchor with only MC items, a mixed-format anchor test containing both MC and CR items; a mixed-format anchor test incorporating common CR item rescoring; and an equivalent groups (EG) design with CR item rescoring, thereby avoiding the need for an anchor test. Designs using either MC items alone or a mixed anchor without CR item rescoring resulted in much larger bias than the other two designs. The EG design with trend scoring resulted in the smallest bias, leading to the smallest root mean squared error value.
Citation	Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, *47*(1), 36–53. https://doi.org/10.1111/j.1745-3984.2009.00098.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2009.00098.x

Title	A Comparison of Chained Linear and Poststratification Linear Equating Under Different Testing Conditions
Abstract	In this study I compared results of chained linear, Tucker, and Levine-observed score equatings under conditions where the new and old forms samples were similar in ability and also when they were different in ability. The length of the anchor test was also varied to examine its effect on the three different equating methods. The three equating methods were compared to a criterion equating to obtain estimates of random equating error, bias, and root mean squared error (RMSE). Results showed that, for most studied conditions, chained linear equating produced fairly good equating results in terms of low bias and RMSE. Levine equating also produced low bias and RMSE in some conditions. Although the Tucker method always produced the lowest random equating error, it produced a larger bias and RMSE than either of the other equating methods. As noted in the literature, these results also suggest that either chained linear or Levine equating be used when new and old form samples differ on ability and/or when the anchor-to-total correlation is not very high. Finally, by testing the missing data assumptions of the three equating methods, this study also shows empirically why an equating method is more or less accurate under certain conditions.
Citation	Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement*, *47*(1), 54–75. https://doi.org/10.1111/j.1745-3984.2009.00099.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2009.00099.x

Title	The Effects of Selection Strategies for Bivariate Loglinear Smoothing Models on NEAT Equating Functions
Abstract	In this study, eight statistical strategies were evaluated for selecting the parameterizations of loglinear models for smoothing the bivariate test score distributions used in nonequivalent groups with anchor test (NEAT) equating. Four of the strategies were based on significance tests of chi-square statistics (Likelihood Ratio, Pearson, Freeman-Tukey, and Cressie-Read) and four additional strategies were based on different evaluations of the Likelihood Ratio Chi-Square statistic (Akaike Information Criterion, Bayesian Information Criterion, Consistent Akaike Information Criterion, and an index traced to Goodman). The focus was the implications of the selection strategies' selection tendencies for the accuracy of chained and poststratification equating functions. The results differentiated the strategies in terms of their tendencies to select models with particular bivariate parameterizations and the implications of these tendencies for equating bias and variability.
Citation	Moses, T., & Holland, P. W. (2010). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement*, *47*(1), 76–91. https://doi.org/10.1111/j.1745-3984.2009.00100.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2009.00100.x

Title	Linking Response-Time Parameters onto a Common Scale
Abstract	Although response times on test items are recorded on a natural scale, the scale for some of the parameters in the lognormal response-time model (van der Linden, 2006) is not fixed. As a result, when the model is used to periodically calibrate new items in a testing program, the parameter are not automatically mapped onto a common scale. Several combinations of linking designs and procedures for the lognormal model are examined that do map parameter estimates onto a common scale. For each of the designs, the standard error of linking is derived. The results are illustrated using examples with simulated data.
Citation	Van Der Linden, W. J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement*, *47*(1), 92–114. https://doi.org/10.1111/j.1745-3984.2009.00101.
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2009.00101.x

Title	Random-Groups Equating with Samples of 50 to 400 Test Takers
Abstract	Five methods for equating in a random groups design were investigated in a series of resampling studies with samples of 400, 200, 100, and 50 test takers. Six operational test forms, each taken by 9,000 or more test takers, were used as item pools to construct pairs of forms to be equated. The criterion equating was the direct equipercentile equating in the group of all test takers. Equating accuracy was indicated by the root-mean-squared deviation, over 1,000 replications, of the sample equatings from the criterion equating. The methods investigated were equipercentile equating of smoothed distributions, linear equating, mean equating, symmetric circle-arc equating, and simplified circle-arc equating. The circle-arc methods produced the most accurate results for all sample sizes investigated, particularly in the upper half of the score distribution. The difference in equating accuracy between the two circle-arc methods was negligible.
Citation	Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*, *47*(2), 175–185. https://doi.org/10.1111/j.1745-3984.2010.00107.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2010.00107.x

Title	Investigating the Effectiveness of Equating Designs for Constructed-Response Tests in Large-Scale Assessments
Abstract	Using data from a large-scale exam, in this study we compared various designs for equating constructed-response (CR) tests to determine which design was most effective in producing equivalent scores across the two tests to be equated. In the context of classical equating methods, four linking designs were examined: (a) an anchor set containing common CR items, (b) an anchor set incorporating common CR items rescored, (c) an external multiple-choice (MC) anchor test, and (d) an equivalent groups design incorporating rescored CR items (no anchor test). The use of CR items without rescoring resulted in much larger bias than the other designs. The use of an external MC anchor resulted in the next largest bias. The use of a rescored CR anchor and the equivalent groups design led to similar levels of equating error.
Citation	Kim, S., Walker, M. E., & McHale, F. (2010). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *Journal of Educational Measurement*, *47*(2), 186–201. https://doi.org/10.1111/j.1745-3984.2010.00108.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2010.00108.x

Title	A New Approach to Comparing Several Equating Methods in the Context of the NEAT Design
Abstract	The nonequivalent groups with anchor test (NEAT) design involves missing data that are missing by design. Three equating methods that can be used with a NEAT design are the frequency estimation equipercentile equating method, the chain equipercentile equating method, and the item-response-theory observed-score-equating method. We suggest an approach to perform a fair comparison of the three methods. The approach is then applied to compare the three equating methods using three data sets from operational tests. For each data set, we examine how the three equating methods perform when the missing data satisfy the assumptions made by only one of these equating methods. The chain equipercentile equating method is somewhat more satisfactory overall than the other methods.
Citation	Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, *47*(3), 261–285. https://doi.org/10.1111/j.1745-3984.2010.00113.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2010.00113.x

Title	Comparisons among Small Sample Equating Methods in a Common-Item Design
Abstract	Score equating based on small samples of examinees is often inaccurate for the examinee populations. We conducted a series of resampling studies to investigate the accuracy of five methods of equating in a common-item design. The methods were chained equipercentile equating of smoothed distributions, chained linear equating, chained mean equating, the symmetric circle-arc method, and the simplified circle-arc method. Four operational test forms, each containing at least 110 items, were used for the equating, with new-form samples of 100, 50, 25, and 10 examinees and reference-form samples three times as large. Accuracy was described in terms of the root-mean-squared difference (over 1,000 replications) of the sample equatings from the criterion equating. Overall, chained mean equating produced the most accurate results for low scores, but the two circle-arc methods produced the most accurate results, particularly in the upper half of the score distribution. The difference in equating accuracy between the two circle-arc methods was negligible.
Citation	Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, *47*(3), 286–298. https://doi.org/10.1111/j.1745-3984.2010.00114.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2010.00114.x

Title	Observed Score Equating Using Discrete and Passage-Based Anchor Items
Abstract	Equating of tests composed of both discrete and passage-based multiple choice items using the nonequivalent groups with anchor test design is popular in practice. In this study, we compared the effect of discrete and passage-based anchor items on observed score equating via simulation. Results suggested that an anchor with a larger proportion of passage-based items, more items in each passage, and/or a larger degree of local dependence among items within one passage produces larger equating errors, especially when the groups taking the new form and the reference form differ in ability. Our findings challenge the common belief that an anchor should be a miniature version of the tests to be equated. Suggestions to practitioners regarding anchor design are also given.
Citation	Zu, J., & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement*, *47*(4), 395–412. https://doi.org/10.1111/j.1745-3984.2010.00120.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2010.00120.x

Title	When Does Scale Anchoring Work? A Case Study
Abstract	Providing information to test takers and test score users about the abilities of test takers at different score levels has been a persistent problem in educational and psychological measurement. Scale anchoring, a technique which describes what students at different points on a score scale know and can do, is a tool to provide such information. Scale anchoring for a test involves a substantial amount of work, both by the statistical analysts and test developers involved with the test. In addition, scale anchoring involves considerable use of subjective judgment, so its conclusions may be questionable. We describe statistical procedures that can be used to determine if scale anchoring is likely to be successful for a test. If these procedures indicate that scale anchoring is unlikely to be successful, then there is little reason to perform a detailed scale anchoring study. The procedures are applied to several data sets from a teachers' licensing test.
Citation	Sinharay, S., Haberman, S. J., & Lee, Y. H. (2011). When does scale anchoring work? A case study. *Journal of Educational Measurement*, *48*(1), 61–80. https://doi.org/10.1111/j.1745-3984.2011.00131.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00131.x

Title	Equating of Augmented Subscores
Abstract	Recently, there has been an increasing level of interest in subscores for their potential diagnostic value. Haberman (2008b) suggested reporting an augmented subscore that is a linear combination of a subscore and the total score. Sinharay and Haberman (2008) and Sinharay (2010) showed that augmented subscores often lead to more accurate diagnostic information than subscores. In order to report augmented subscores operationally, they should be comparable across the different forms of a test. One way to achieve comparability is to equate them. We suggest several methods for equating augmented subscores. Results from several operational and simulated data sets show that the error in the equating of augmented subscores appears to be small in most practical situations.
Citation	Sinharay, S., & Haberman, S. J. (2011). Equating of augmented subscores. *Journal of Educational Measurement*, *48*(2), 122–145. https://doi.org/10.1111/j.1745-3984.2011.00137.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00137.x

Title	Local Linear Observed-Score Equating
Abstract	Two methods of local linear observed-score equating for use with anchor-test and single-group designs are introduced. In an empirical study, the two methods were compared with the current traditional linear methods for observed-score equating. As a criterion, the bias in the equated scores relative to true equating based on Lord's (1980) definition of equity was used. The local method for the anchor-test design yielded minimum bias, even for considerable variation of the relative difficulties of the two test forms and the length of the anchor test. Among the traditional methods, the method of chain equating performed best. The local method for single-group designs yielded equated scores with bias comparable to the traditional methods. This method, however, appears to be of theoretical interest because it forces us to rethink the relationship between score equating and regression.
Citation	Wiberg, M., & van der Linden, W. J. (2011). Local linear observed-score equating. *Journal of educational measurement*, *48*(3), 229–254. https://doi.org/10.1111/j.1745-3984.2011.00148.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00148.x

Title	Futility of Log-Linear Smoothing when Equating with Unrepresentative Small Samples
Abstract	The impact of log-linear presmoothing on the accuracy of small sample chained equipercentile equating was evaluated under two conditions. In the first condition the small samples differed randomly in ability from the target population. In the second condition the small samples were systematically different from the target population. Results showed that equating with small samples (e.g., N < 25 or 50) using either raw or smoothed score distributions led to considerable large random equating error (although smoothing reduced random equating error). Moreover, when the small samples were not representative of the target population, the amount of equating bias also was quite large. It is concluded that although presmoothing can reduce random equating error, it is not likely to reduce equating bias caused by using an unrepresentative sample. Other alternatives to the small sample equating problem (e.g., the SiGNET design) which focus more on improving data collection are discussed.
Citation	Puhan, G. (2011). Futility of log-linear smoothing when equating with unrepresentative small samples. *Journal of Educational Measurement*, *48*(3), 274–292. https://doi.org/10.1111/j.1745-3984.2011.00147.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00147.x

Title	Test Score Equating Using a Mini-Version Anchor and a Midi Anchor: A Case Study Using SAT® Data
Abstract	This study explores an anchor that is different from the traditional miniature anchor in test score equating. In contrast to a traditional "mini" anchor that has the same spread of item difficulties as the tests to be equated, the studied anchor, referred to as a "midi" anchor (Sinharay & Holland), has a smaller spread of item difficulties than the tests to be equated. Both anchors were administered in an operational SAT administration and the impact of anchor type on equating was evaluated with respect to systematic error or equating bias. Contradicting the popular belief that the mini anchor is best, the results showed that the mini anchor does not always produce more accurate equating functions than the midi anchor; the midi anchor was found to perform as well as or even better than the mini anchor. Because testing programs usually have more middle difficulty items and few very hard or very easy items, midi external anchors are operationally easier to build. Therefore, the results of our study provide evidence in favor of the midi anchor, the use of which will lead to cost saving with no reduction in equating quality.
Citation	Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT® data. *Journal of Educational Measurement*, *48*(4), 361–379. https://doi.org/10.1111/j.1745-3984.2011.00150.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00150.x

Title	Observed Score Linear Equating with Covariates
Abstract	This paper examined observed score linear equating in two different data collection designs, the equivalent groups design and the nonequivalent groups design, when information from covariates (i.e., background variables correlated with the test scores) was included. The main purpose of the study was to examine the effect (i.e., bias, variance, and mean squared error) on the estimators of including this additional information. A model for observed score linear equating with covariates first was suggested. As a second step, the model was used in a simulation study to show that the use of covariates such as gender and education can increase the accuracy of an equating by reducing the mean squared error of the estimators. Finally, data from two administrations of the Swedish Scholastic Assessment Test were used to illustrate the use of the model.
Citation	Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, *48*(4), 419–440. https://doi.org/10.1111/j.1745-3984.2011.00153.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00153.x

Title	Addressing the Extreme Assumptions of Presumed Linkings
Abstract	The interpretability of score comparisons depends on the design and execution of a sound data collection plan and the establishment of linkings between these scores. When comparisons are made between scores from two or more assessments that are built to different specifications and are administered to different populations under different conditions, the validity of the comparisons hinges on untestable assumptions. For example, tests administered across different disability groups or tests administered to different language groups produce scores for which implicit linkings are presumed to hold. Presumed linking makes use of extreme assumptions to produce links between scores on tests in the absence of common test material or equivalent groups of test takers. These presumed linkings lead to dubious interpretations. This article suggests an approach that indirectly assesses the validity of these presumed linkings among scores on assessments that contain neither equivalent groups nor common anchor material.
Citation	Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement*, *49*(1), 1–18. https://doi.org/10.1111/j.1745-3984.2011.00157.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00157.x

Title	Evaluating Equating Results: Percent Relative Error for Chained Kernel Equating
Abstract	This article presents a method for evaluating equating results. Within the kernel equating framework, the percent relative error (PRE) for chained equipercentile equating was computed under the nonequivalent groups with anchor test (NEAT) design. The method was applied to two data sets to obtain the PRE, which can be used to measure equating effectiveness. The study compared the PRE results for chained and poststratification equating. The results indicated that the chained method transformed the new form score distribution to the reference form scale more effectively than the poststratification method. In addition, the study found that in chained equating, the population weight had impact on score distributions over the target population but not on the equating and PRE results.
Citation	Jiang, Y., von Davier, A. A., & Chen, H. (2012). Evaluating equating results: Percent relative error for chained kernel equating. *Journal of Educational Measurement*, *49*(1), 39–58. https://doi.org/10.1111/j.1745-3984.2011.00159.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00159.x

Title	Scaling, Linking, and Reporting in a Periodic Assessment System
Abstract	A new entry in the testing lexicon is through-course summative assessment, a system consisting of components administered periodically during the academic year. As defined in the Race to the Top program, these assessments are intended to yield a yearly summative score for accountability purposes. They must provide for both individual and group proficiency estimates and allow for the measurement of growth. They must accommodate students who vary in their patterns of curricular exposure. Because they are meant to provide actionable information to teachers they must be instructionally sensitive, so item-operating characteristics can be expected to change relative to one another as a function of patterns of curricular exposure. This paper discusses methodology one can draw upon to tackle this ambitious collection of inferences. We consider a modeling framework that consists of an item response theory component and a population component, as in the National Assessment of Educational Progress, and show how performance and growth could be expressed in terms of expected performance on a market basket of tasks. We discuss conditions under which modeling simplifications might be possible and discuss studies that would be needed to fit models, estimate parameters, and evaluate data requirements.
Citation	Mislevy, R. J., & Zwick, R. (2012). Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement*, *49*(2), 148–166. https://doi.org/10.1111/j.1745-3984.2012.00166.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2012.00166.x

Title	The Accuracy and Consistency of a Series of IRT True Score Equatings
Abstract	This study investigates a sequence of item response theory (IRT) true score equatings based on various scale transformation approaches and evaluates equating accuracy and consistency over time. The results show that the biases and sample variances for the IRT true score equating (both direct and indirect) are quite small (except for the mean/sigma method). The biases and sample variances for the equating functions based on the characteristic curve methods and concurrent calibrations for adjacent forms are smaller than the biases and variances for the equating functions based on the moment methods. In addition, the IRT true score equating is also compared to the chained equipercentile equating, and we observe that the sample variances for the chained equipercentile equating are much smaller than the variances for the IRT true score equating with an exception at the low scores.
Citation	Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement*, *49*(2), 167–189. https://doi.org/10.1111/j.1745-3984.2012.00167.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2012.00167.x

Title	Standard Error of Linear Observed-Score Equating for the NEAT Design With Nonnormally Distributed Data
Abstract	In the nonequivalent groups with anchor test (NEAT) design, the standard error of linear observed-score equating is commonly estimated by an estimator derived assuming multivariate normality. However, real data are seldom normally distributed, causing this normal estimator to be inconsistent. A general estimator, which does not rely on the normality assumption, would be preferred, because it is asymptotically accurate regardless of the distribution of the data. In this article, an analytical formula for the standard error of linear observed-score equating, which characterizes the effect of nonnormality, is obtained under elliptical distributions. Using three large-scale real data sets as the populations, resampling studies are conducted to empirically evaluate the normal and general estimators of the standard error of linear observed-score equating. The effect of sample size (50, 100, 250, or 500) and equating method (chained linear, Tucker, or Levine observed-score equating) are examined. Results suggest that the general estimator has smaller bias than the normal estimator in all 36 conditions; it has larger standard error when the sample size is at least 100; and it has smaller root mean squared error in all but one condition. An R program is also provided to facilitate the use of the general estimator.
Citation	Zu, J., & Yuan, K. H. (2012). Standard error of linear observed-score equating for the NEAT design with nonnormally distributed data. *Journal of Educational Measurement*, *49*(2), 190–213.
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2012.00168.x

Title	A Comparison Between Linear IRT Observed-Score Equating and Levine Observed-Score Equating Under the Generalized Kernel Equating Framework
Abstract	In this article, linear item response theory (IRT) observed-score equating is compared under a generalized kernel equating framework with Levine observed-score equating for nonequivalent groups with anchor test design. Interestingly, these two equating methods are closely related despite being based on different methodologies. Specifically, when using data from IRT models, linear IRT observed-score equating is virtually identical to Levine observed-score equating. This leads to the conclusion that poststratification equating based on true anchor scores can be viewed as the curvilinear Levine observed-score equating.
Citation	Chen, H. (2012). A comparison between linear IRT observed-score equating and Levine observed-score equating under the generalized kernel equating framework. *Journal of Educational Measurement*, *49*(3), 269–284. https://doi.org/10.1111/j.1745-3984.2012.00175.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2012.00175.x

Title	Choosing Among Tucker or Chained Linear Equating in Two Testing Situations: Rater Comparability Scoring and Randomly Equivalent Groups With an Anchor
Abstract	Tucker and chained linear equatings were evaluated in two testing scenarios. In Scenario 1, referred to as rater comparability scoring and equating, the anchor-to-total correlation is often very high for the new form but moderate for the reference form. This may adversely affect the results of Tucker equating, especially if the new and reference form samples differ in ability. In Scenario 2, the new and reference form samples are randomly equivalent but the correlation between the anchor and total scores is low. When the correlation between the anchor and total scores is low, Tucker equating assumes that the new and reference form samples are similar in ability (which, with randomly equivalents groups, is the correct assumption). Thus Tucker equating should produce accurate results. Results indicated that in Scenario 1, the Tucker results were less accurate than the chained linear equating results. However, in Scenario 2, the Tucker results were more accurate than the chained linear equating results. Some implications are discussed.
Citation	Puhan, G. (2012). Choosing among Tucker or chained linear equating in two testing situations: Rater comparability scoring and randomly equivalent groups with an anchor. *Journal of Educational Measurement*, *49*(3), 312–329. https://doi.org/10.1111/j.1745-3984.2012.00177.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2012.00177.x

Title	Comparison of the One- and Bi-Direction Chained Equipercentile Equating
Abstract	This study investigated differences between two approaches to chained equipercentile (CE) equating (one- and bi-direction CE equating) in nearly equal groups and relatively unequal groups. In one-direction CE equating, the new form is linked to the anchor in one sample of examinees and the anchor is linked to the reference form in the other sample. In bi-direction CE equating, the anchor is linked to the new form in one sample of examinees and to the reference form in the other sample. The two approaches were evaluated in comparison to a criterion equating function (i.e., equivalent groups equating) using indexes such as root expected squared difference, bias, standard error of equating, root mean squared error, and number of gaps and bumps. The overall results across the equating situations suggested that the two CE equating approaches produced very similar results, whereas the bi-direction results were slightly less erratic, smoother (i.e., fewer gaps and bumps), usually closer to the criterion function, and also less variable.
Citation	Oh, H., & Moses, T. (2012). Comparison of the one- and bi-direction chained equipercentile equating. *Journal of Educational Measurement*, *49*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2012.00183.x
Link	https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2012.00183.x

Title	Measuring Growth With Vertical Scales
Abstract	A vertical score scale is needed to measure growth across multiple tests in terms of absolute changes in magnitude. Since the warrant for subsequent growth interpretations depends upon the assumption that the scale has interval properties, the validation of a vertical scale would seem to require methods for distinguishing interval scales from ordinal scales. In taking up this issue, two different perspectives on educational measurement are contrasted: a metaphorical perspective and a classical perspective. Although the metaphorical perspective is more predominant, at present it provides no objective methods whereby the properties of a vertical scale can be validated. In contrast, when taking a classical perspective, the axioms of additive conjoint measurement can be used to test the hypothesis that the latent variable underlying a vertical scale is quantitative (supporting ratio or interval properties) rather than merely qualitative (supporting ordinal or nominal properties). The application of such an approach is illustrated with both a hypothetical example and by drawing upon recent research that has been conducted on the Lexile scale for reading comprehension.
Citation	Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204–226. https://doi.org/10.1111/jedm.12011
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12011

Title	Some Conceptual Issues in Observed-Score Equating
Abstract	In spite of all of the technical progress in observed-score equating, several of the more conceptual aspects of the process still are not well understood. As a result, the equating literature struggles with rather complex criteria of equating, lack of a test-theoretic foundation, confusing terminology, and ad hoc analyses. A return to Lord's foundational criterion of equity of equating, a derivation of the true equating transformation from it, and mainstream statistical treatment of the problem of estimating the transformation for various data-collection designs exist as a solution to the problem.
Citation	van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement*, *50*(3), 249–285. https://doi.org/10.1111/jedm.12014
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12014

Title	Comments on van der Linden's Critique and Proposal for Equating
Abstract	While agreeing with van der Linden (this issue) that test equating needs better theoretical underpinnings, my comments criticize several aspects of his article. His examples are, for the most part, worthless; he does not use well-established terminology correctly; his view of 100 years of attempts to give a theoretical basis for equating is unreasonably dismissive; he exhibits no understanding of the role of the synthetic population for anchor test equating for the nonequivalent groups with anchor test design; he is obtuse regarding the condition of symmetry, requiring it of the estimand but not of the estimator; and his proposal for a foundational basis for all test equating, the "true equating transformation," allows a different equating function for every examinee, which is way past what equating actually does or hopes to achieve. Most importantly, he appears to think that criticism of others is more important than improved insight that moves a field forward based on the work of many other theorists whose contributions have improved the practice of equating.
Citation	Holland, P. W. (2013). Comments on van der Linden's critique and proposal for equating. *Journal of Educational Measurement*, *50*(3), 286–294. https://doi.org/10.1111/jedm.12015
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12015

Title	Local Equating Using the Rasch Model, the OPLM, and the 2PL IRT Model—or—What Is It Anyway if the Model Captures Everything There Is to Know About the Test Takers?
Abstract	Local equating (LE) is based on Lord's criterion of equity. It defines a family of true transformations that aim at the ideal of equitable equating. van der Linden (this issue) offers a detailed discussion of common issues in observed-score equating relative to this local approach. By assuming an underlying item response theory model, one of the main features of LE is that it adjusts the equated raw scores using conditional distributions of raw scores given an estimate of the ability of interest. In this article, we argue that this feature disappears when using a Rasch model for the estimation of the true transformation, while the one-parameter logistic model and the two-parameter logistic model do provide a local adjustment of the equated score.
Citation	von Davier, M., & von Davier, A. A. (2013). Local equating using the Rasch model, the OPLM, and the 2PL IRT model—or—what is it anyway if the model captures everything there is to know about the test takers?. *Journal of Educational Measurement*, *50*(3), 295–303. https://doi.org/10.1111/jedm.12016
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12016

Title	On Attempting to Do What Lord Said Was Impossible: Commentary on van der Linden's "Some Conceptual Issues in Observed-Score Equating"
Abstract	van der Linden (this issue) uses words differently than Holland and Dorans. This difference in language usage is a source of some confusion in van der Linden's critique of what he calls equipercentile equating. I address these differences in language. van der Linden maintains that there are only two requirements for score equating. I maintain that the requirements he discards have practical utility and are testable. The score equity requirement proposed by Lord suggests that observed score equating was either unnecessary or impossible. Strong equity serves as the fulcrum for van der Linden's thesis. His proposed solution to the equity problem takes inequitable measures and aligns conditional error score distributions, resulting in a family of linking functions, one for each level of θ. In reality, θ is never known. Use of an anchor test as a proxy poses many practical problems, including defensibility.
Citation	Dorans, N. J. (2013). On attempting to do what Lord said was impossible: Commentary on van der Linden's "some conceptual issues in observed-score equating". *Journal of Educational Measurement*, *50*(3), 304–314. https://doi.org/10.1111/jedm.12017
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12017

Title	Statistical Models and Inference for the True Equating Transformation in the Context of Local Equating
Abstract	Based on Lord's criterion of equity of equating, van der Linden (this issue) revisits the so-called local equating method and offers alternative as well as new thoughts on several topics including the types of transformations, symmetry, reliability, and population invariance appropriate for equating. A remarkable aspect is to define equating as a standard statistical inference problem in which the true equating transformation is the parameter of interest that has to be estimated and assessed as any standard evaluation of an estimator of an unknown parameter in statistics. We believe that putting equating methods in a general statistical model framework would be an interesting and useful next step in the area. van der Linden's conceptual article on equating is certainly an important contribution to this task.
Citation	Jorge, G. B., & von Davier, M. (2013). Statistical models and inference for the true equating transformation in the context of local equating. *Journal of Educational Measurement*, *50*(3), 315–320. https://doi.org/10.1111/jedm.12018
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12018

Title	Comments on "Some Conceptual Issues in Observed-Score Equating" by Wim J. van der Linden
Abstract	The van der Linden article (this issue) provides a roadmap for future research in equating. My belief is that the roadmap begins and ends with collecting auxiliary data that can be utilized to provide improved equating, especially when data are sparse or equating beyond simple moments is desired.
Citation	Bradlow, E. T. (2013). Comments on "some conceptual issues in observed-score equating" by Wim J. van der Linden. *Journal of Educational Measurement*, *50*(3), 321–323. https://doi.org/10.1111/jedm.12019
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12019

Title	More Issues in Observed-Score Equating
Abstract	This article is a response to the commentaries on the position paper on observed-score equating by van der Linden (this issue). The response focuses on the more general issues in these commentaries, such as the nature of the observed scores that are equated, the importance of test-theory assumptions in equating, the necessity to use multiple equating transformations, and the choice of conditioning variables in equating.
Citation	van der Linden, W. J. (2013). More issues in observed-score equating. *Journal of Educational Measurement*, *50*(3), 324–337. https://doi.org/10.1111/jedm.12020
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12020

Title	Situations Where It Is Appropriate to Use Frequency Estimation Equipercentile Equating
Abstract	In operational equating situations, frequency estimation equipercentile equating is considered only when the old and new groups have similar abilities. The frequency estimation assumptions are investigated in this study under various situations from both the levels of theoretical interest and practical use. It shows that frequency estimation equating can be used under circumstances when it is not normally used. To link theoretical results with practice, statistical methods are proposed for checking frequency estimation assumptions based on available data: observed-score distributions and item difficulty distributions of the forms. In addition to the conventional use of frequency estimation equating when the group abilities are similar, three situations are identified when the group abilities are dissimilar: (a) when the two forms and the observed conditional score distributions are similar the two forms and the observed conditional score distributions are similar (in this situation, the frequency estimation equating assumptions are likely to hold, and frequency estimation equating is appropriate); (b) when forms are similar but the observed conditional score distributions are not (in this situation, frequency estimation equating is not appropriate); and (c) when forms are not similar but the observed conditional score distributions are (frequency estimation equating is not appropriate). Statistical analysis procedures for comparing distributions are provided. Data from a large-scale test are used to illustrate the use of frequency estimation equating when the group difference in ability is large.
Citation	Guo, H., Oh, H. J., & Eignor, D. (2013). Situations where it is appropriate to use frequency estimation equipercentile equating. *Journal of Educational Measurement*, *50*(3), 338–354. https://doi.org/10.1111/jedm.12021
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12021

Title	The Long-Term Sustainability of IRT Scaling Methods in Mixed-Format Tests
Abstract	Due to recent research in equating methodologies indicating that some methods may be more susceptible to the accumulation of equating error over multiple administrations, the sustainability of several item response theory methods of equating over time was investigated. In particular, the paper is focused on two equating methodologies: fixed common item parameter scaling (with two variations, FCIP-1 and FCIP-2) and the Stocking and Lord characteristic curve scaling technique in the presence of nonequivalent groups. Results indicated that the improvements made to fixed common item parameter scaling in the FCIP-2 method were sustained over time. FCIP-2 and Stocking and Lord characteristic curve scaling performed similarly in many instances and produced more accurate results than FCIP-1. The relative performance of FCIP-2 and Stocking and Lord characteristic curve scaling depended on the nature of the change in the ability distribution: Stocking and Lord characteristic curve scaling captured the change in the distribution more accurately than FCIP-2 when the change was different across the ability distribution; FCIP-2 captured the changes more accurately when the change was consistent across the ability distribution.
Citation	Keller, L. A., & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement*, *50*(4), 390–407. https://doi.org/10.1111/jedm.12025
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12025

Title	Adjoined Piecewise Linear Approximations (APLAs) for Equating: Accuracy Evaluations of a Postsmoothing Equating Method
Abstract	The purpose of this study was to evaluate the use of adjoined and piecewise linear approximations (APLAs) of raw equipercentile equating functions as a postsmoothing equating method. APLAs are less familiar than other postsmoothing equating methods (i.e., cubic splines), but their use has been described in historical equating practices of large-scale testing programs. This study used simulations to evaluate APLA equating results and compare these results with those from cubic spline postsmoothing and from several presmoothing equating methods. The overall results suggested that APLAs based on four line segments have accuracy advantages similar to or better than cubic splines and can sometimes produce more accurate smoothed equating functions than those produced using presmoothing methods.
Citation	Moses, T. (2013). Adjoined piecewise linear approximations (APLAs) for equating: Accuracy evaluations of a postsmoothing equating method. *Journal of Educational Measurement*, *50*(4), 427–446. https://doi.org/10.1111/jedm.12027
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12027

Title	IRT Model Misspecification and Measurement of Growth in Vertical Scaling
Abstract	Functional form misfit is frequently a concern in item response theory (IRT), although the practical implications of misfit are often difficult to evaluate. In this article, we illustrate how seemingly negligible amounts of functional form misfit, when systematic, can be associated with significant distortions of the score metric in vertical scaling contexts. Our analysis uses two- and three-parameter versions of Samejima's logistic positive exponent model (LPE) as a data generating model. Consistent with prior work, we find LPEs generally provide a better comparative fit to real item response data than traditional IRT models (2PL, 3PL). Further, our simulation results illustrate how 2PL- or 3PL-based vertical scaling in the presence of LPE-induced misspecification leads to an artificial growth deceleration across grades, consistent with that commonly seen in vertical scaling studies. The results raise further concerns about the use of standard IRT models in measuring growth, even apart from the frequently cited concerns of construct shift/multidimensionality across grades.
Citation	Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, *51*(2), 141–162. https://doi.org/10.1111/jedm.12039
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12039

Title	Optimal Bandwidth Selection in Observed-Score Kernel Equating
Abstract	The selection of bandwidth in kernel equating is important because it has a direct impact on the equated test scores. The aim of this article is to examine the use of double smoothing when selecting bandwidths in kernel equating and to compare double smoothing with the commonly used penalty method. This comparison was made using both an equivalent groups design and a nonequivalent group with anchor test design. The performance of the methods was evaluated through simulation studies using both symmetric and skewed score distributions. In addition, the bandwidth selection methods were applied to real data from a college admissions test. The results show that the traditional penalty method works well although double smoothing is a viable alternative because it performs reasonably well compared to the traditional method.
Citation	Häggström, J., & Wiberg, M. (2014). Optimal bandwidth selection in observed-score kernel equating. *Journal of Educational Measurement*, *51*(2), 201–211. https://doi.org/10.1111/jedm.12042
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12042

Title	Improving the Bandwidth Selection in Kernel Equating
Abstract	We investigate the current bandwidth selection methods in kernel equating and propose a method based on Silverman's rule of thumb for selecting the bandwidth parameters. In kernel equating, the bandwidth parameters have previously been obtained by minimizing a penalty function. This minimization process has been criticized by practitioners for being too complex and that it does not offer sufficient smoothing in certain cases. In addition, the bandwidth parameters have been treated as constants in the derivation of the standard error of equating even when they were selected by considering the observed data. Here, the bandwidth selection is simplified, and modified standard errors of equating (SEEs) that reflect the bandwidth selection method are derived. The method is illustrated with real data examples and simulated data.
Citation	Andersson, B., & von Davier, A. A. (2014). Improving the bandwidth selection in kernel equating. *Journal of Educational Measurement*, *51*(3), 223–238. https://doi.org/10.1111/jedm.12044
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12044

Title	Preequating With Empirical Item Characteristic Curves: An Observed-Score Preequating Method
Abstract	Preequating is in demand because it reduces score reporting time. In this article, we evaluated an observed-score preequating method: the empirical item characteristic curve (EICC) method, which makes preequating without item response theory (IRT) possible. EICC preequating results were compared with a criterion equating and with IRT true-score preequating conversions. Results suggested that the EICC preequating method worked well under the conditions considered in this study. The difference between the EICC preequating conversion and the criterion equating was smaller than .5 raw-score points (a practical criterion often used to evaluate equating quality) between the 5th and 95th percentiles of the new form total score distribution. EICC preequating also performed similarly or slightly better than IRT true-score preequating.
Citation	Zu, J., & Puhan, G. (2014). Preequating with empirical item characteristic curves: An observed-score preequating method. *Journal of Educational Measurement*, *51*(3), 281–300. https://doi.org/10.1111/jedm.12047
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12047

Title	Section Preequating Under the Equivalent Groups Design Without IRT
Abstract	In this article, we introduce a section preequating (SPE) method (linear and nonlinear) under the randomly equivalent groups design. In this equating design, sections of Test X (a future new form) and another existing Test Y (an old form already on scale) are administered. The sections of Test X are equated to Test Y, after adjusting for the imperfect correlation between sections of Test X, to obtain the equated score on the complete form of X. Simulations and a real-data application show that the proposed SPE method is fairly simple and accurate.
Citation	Guo, H., & Puhan, G. (2014). Section preequating under the equivalent groups design without IRT. *Journal of Educational Measurement*, *51*(3), 301–317. https://doi.org/10.1111/jedm.12049
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12049

Title	A General Linear Method for Equating With Small Samples
Abstract	Research on equating with small samples has shown that methods with stronger assumptions and fewer statistical estimates can lead to decreased error in the estimated equating function. This article introduces a new approach to linear observed-score equating, one which provides flexible control over how form difficulty is assumed versus estimated to change across the score scale. A general linear method is presented as an extension of traditional linear methods. The general method is then compared to other linear and nonlinear methods in terms of accuracy in estimating a criterion equating function. Results from two parametric bootstrapping studies based on real data demonstrate the usefulness of the general linear method.
Citation	Albano, A. D. (2015). A general linear method for equating with small samples. *Journal of Educational Measurement*, *52*(1), 55–69. https://doi.org/10.1111/jedm.12062
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12062

Title	Asymptotic Standard Errors for Item Response Theory True Score Equating of Polytomous Items
Abstract	Building on previous works by Lord and Ogasawara for dichotomous items, this article proposes an approach to derive the asymptotic standard errors of item response theory true score equating involving polytomous items, for equivalent and nonequivalent groups of examinees. This analytical approach could be used in place of empirical methods like the bootstrap method, to obtain standard errors of equated scores. Formulas are introduced to obtain the derivatives for computing the asymptotic standard errors. The approach was validated using mean-mean, mean-sigma, random-groups, or concurrent calibration equating of simulated samples, for tests modeled using the generalized partial credit model or the graded response model.
Citation	Cher Wong, C. (2015). Asymptotic standard errors for item response theory true score equating of polytomous items. *Journal of Educational Measurement*, *52*(1), 106–120. https://doi.org/10.1111/jedm.12065
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12065

Title	The Effect of Differential Motivation on IRT Linking
Abstract	The purpose of this study was to investigate whether simulated differential motivation between the stakes for operational tests and anchor items produces an invalid linking result if the Rasch model is used to link the operational tests. This was done for an external anchor design and a variation of a pretest design. The study also investigated whether a constrained mixture Rasch model could identify latent classes in such a way that one latent class represented high-stakes responding while the other represented low-stakes responding. The results indicated that for an external anchor design, the Rasch linking result was only biased when the motivation level differed between the subpopulations to which the anchor items were administered. However, the mixture Rasch model did not identify the classes representing low-stakes and high-stakes responding. When a pretest design was used to link the operational tests by means of a Rasch model, the linking result was found to be biased in each condition. Bias increased as percentage of students showing low-stakes responding to the anchor items increased. The mixture Rasch model only identified the classes representing low-stakes and high-stakes responding under a limited number of conditions.
Citation	Mittelhaëuser, M. A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, *52*(3), 339–358. https://doi.org/10.1111/jedm.12080
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12080

Title	Statistical Assessment of Estimated Transformations in Observed-Score Equating
Abstract	Equating methods make use of an appropriate transformation function to map the scores of one test form into the scale of another so that scores are comparable and can be used interchangeably. The equating literature shows that the ways of judging the success of an equating (i.e., the score transformation) might differ depending on the adopted framework. Rather than targeting different parts of the equating process and aiming to evaluate the process from different aspects, this article views the equating transformation as a standard statistical estimator and discusses how this estimator should be assessed in an equating framework. For the kernel equating framework, a numerical illustration shows the potentials of viewing the equating transformation as a statistical estimator as opposed to assessing it using equating-specific criteria. A discussion on how this approach can be used to compare other equating estimators from different frameworks is also included.
Citation	Wiberg, M., & Gonzalez, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement*, *53*(1), 106–125. https://doi.org/10.1111/jedm.12103
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12103

Title	A Comparison of Linking Methods for Estimating National Trends in International Comparative Large-Scale Assessments in the Presence of Cross-National DIF
Abstract	Trend estimation in international comparative large-scale assessments relies on measurement invariance between countries. However, cross-national differential item functioning (DIF) has been repeatedly documented. We ran a simulation study using national item parameters, which required trends to be computed separately for each country, to compare trend estimation performances to two linking methods employing international item parameters across several conditions. The trend estimates based on the national item parameters were more accurate than the trend estimates based on the international item parameters when cross-national DIF was present. Moreover, the use of fixed common item parameter calibrations led to biased trend estimates. The detection and elimination of DIF can reduce this bias but is also likely to increase the total error.
Citation	Sachse, K. A., Roppelt, A., & Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *Journal of Educational Measurement*, *53*(2), 152–171. https://doi.org/10.1111/jedm.12106
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12106

Title	Equating With Miditests Using IRT
Abstract	The equating performance of two internal anchor test structures—miditests and minitests—is studied for four IRT equating methods using simulated data. Originally proposed by Sinharay and Holland, miditests are anchors that have the same mean difficulty as the overall test but less variance in item difficulties. Four popular IRT equating methods were tested, and both the means and SDs of the true ability of the group to be equated were varied. We evaluate equating accuracy marginally and conditional on true ability. Our results suggest miditests perform about as well as traditional minitests for most conditions. Findings are discussed in terms of comparability to the typical minitest design and the trade-off between accuracy and flexibility in test construction.
Citation	Fitzpatrick, J., & Skorupski, W. P. (2016). Equating with miditests using IRT. *Journal of Educational Measurement*, *53*(2), 172–189. https://doi.org/10.1111/jedm.12109
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12109

Title	Asymptotic Standard Errors of Observed-Score Equating With Polytomous IRT Models
Abstract	In observed-score equipercentile equating, the goal is to make scores on two scales or tests measuring the same construct comparable by matching the percentiles of the respective score distributions. If the tests consist of different items with multiple categories for each item, a suitable model for the responses is a polytomous item response theory (IRT) model. The parameters from such a model can be utilized to derive the score probabilities for the tests and these score probabilities may then be used in observed-score equating. In this study, the asymptotic standard errors of observed-score equating using score probability vectors from polytomous IRT models are derived using the delta method. The results are applied to the equivalent groups design and the nonequivalent groups design with either chain equating or poststratification equating within the framework of kernel equating. The derivations are presented in a general form and specific formulas for the graded response model and the generalized partial credit model are provided. The asymptotic standard errors are accurate under several simulation conditions relating to sample size, distributional misspecification and, for the nonequivalent groups design, anchor test length.
Citation	Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT models. *Journal of Educational Measurement*, *53*(4), 459–477. https://doi.org/10.1111/jedm.12126
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12126

Title	Further Study of the Choice of Anchor Tests in Equating
Abstract	In this study, we describe what factors influence the observed score correlation between an (external) anchor test and a total test. We show that the anchor to full-test observed score correlation is based on two components: the true score correlation between the anchor and total test, and the reliability of the anchor test. Findings using an analytical approach suggest that making an anchor test a miditest does not generally maximize the anchor to total test correlation. Results are discussed in the context of what conditions maximize the correlations between the anchor and total test.
Citation	Trierweiler, T. J., Lewis, C., & Smith, R. L. (2016). Further study of the choice of anchor tests in equating. *Journal of Educational Measurement*, *53*(4), 498–518. https://doi.org/10.1111/jedm.12128
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12128

Title	Structural Zeros and Their Implications With Log-Linear Bivariate Presmoothing Under the Internal-Anchor Design
Abstract	In equating, when common items are internal and scoring is conducted in terms of the number of correct items, some pairs of total scores (X) and common-item scores (V) can never be observed in a bivariate distribution of X and V; these pairs are called structural zeros. This simulation study examines how equating results compare for different approaches to handling structural zeros. The study considers four approaches: the no-smoothing, unique-common, total-common, and adjusted total-common approaches. This study led to four main findings: (1) the total-common approach generally had the worst results; (2) for relatively small effect sizes, the unique-common approach generally had the smallest overall error; (3) for relatively large effect sizes, the adjusted total-common approach generally had the smallest overall error; and, (4) if sole interest focuses on reducing bias only, the adjusted total-common approach was generally preferable. These results suggest that, when common items are internal and log-linear bivariate presmoothing is performed, structural zeros should be maintained, even if there is some loss in the moment preservation property.
Citation	Kim, H. J., Brennan, R. L., & Lee, W.-C. (2017). Structural zeros and their implications with log-linear bivariate presmoothing under the internal-anchor design. *Journal of Educational Measurement*, *54*(2), 145–164. https://doi.org/10.1111/jedm.12138
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12138

Title	Stabilizing Conditional Standard Errors of Measurement in Scale Score Transformations
Abstract	The focus of this article is on scale score transformations that can be used to stabilize conditional standard errors of measurement (CSEMs). Three transformations for stabilizing the estimated CSEMs are reviewed, including the traditional arcsine transformation, a recently developed general variance stabilization transformation, and a new method proposed in this article involving cubic transformations. Two examples are provided and the three scale score transformations are compared in terms of how well they stabilize CSEMs estimated from compound binomial and item response theory (IRT) models. Advantages of the cubic transformation are demonstrated with respect to CSEM stabilization and other scaling criteria (e.g., scale score distributions that are more symmetric).
Citation	Moses, T., & Kim, Y. (2017). Stabilizing conditional standard errors of measurement in scale score transformations. *Journal of Educational Measurement*, *54*(2), 184–199. https://doi.org/10.1111/jedm.12140
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12140

Title	Optimal Linking Design for Response Model Parameters
Abstract	Linking functions adjust for differences between identifiability restrictions used in different instances of the estimation of item response model parameters. These adjustments are necessary when results from those instances are to be compared. As linking functions are derived from estimated item response model parameters, parameter estimation error automatically propagates into linking error. This article explores an optimal linking design approach in which mixed-integer programming is used to select linking items to minimize linking error. Results indicate that the method holds promise for selection of linking items.
Citation	Barrett, M. D., & van der Linden, W. J. (2017). Optimal linking design for response model parameters. *Journal of Educational Measurement*, *54*(3), 285–305. https://doi.org/10.1111/jedm.12145
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12145

Title	Gathering and Evaluating Validity Evidence: The Generalized Assessment Alignment Tool
Abstract	Alignment is an essential piece of validity evidence for both educational (K-12) and credentialing (licensure and certification) assessments. In this article, a comprehensive review of commonly used contemporary alignment procedures is provided; some key weaknesses in current alignment approaches are identified; principles for evaluating alignment methods are distilled; and a new approach to investigating alignment is proposed and illustrated. The article concludes with suggestions for alignment research and practice.
Citation	Cizek, G. J., Kosh, A. E., & Toutkoushian, E. K. (2018). Gathering and evaluating validity evidence: The generalized assessment alignment tool. *Journal of Educational Measurement*, *55*(4), 477–512. https://doi.org/10.1111/jedm.12189
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12189

Title	A Comparison of Strategies for Smoothing Parameter Selection for Mixed-Format Tests Under the Random Groups Design
Abstract	Smoothing techniques are designed to improve the accuracy of equating functions. The main purpose of this study is to compare seven model selection strategies for choosing the smoothing parameter (C) for polynomial loglinear presmoothing and one procedure for model selection in cubic spline postsmoothing for mixed-format pseudo tests under the random groups design. These model selection strategies were compared for four sample sizes (500, 1,000, 2,000, and 3,000) and two content areas (Advanced Placement [AP] Biology and AP Environmental Science). For polynomial loglinear presmoothing, the Akaike information criterion (AIC) was the only statistic that reduced both random equating error and total equating error in all investigated conditions. Cubic spline postsmoothing tended to produce more accurate results than any of the model selection strategies in polynomial loglinear smoothing.
Citation	Liu, C., & Kolen, M. J. (2018). A comparison of strategies for smoothing parameter selection for mixed-format tests under the random groups design. *Journal of Educational Measurement*, *55*(4), 564–581. https://doi.org/10.1111/jedm.12192
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12192

Title	Scale Alignment in Between-Item Multidimensional Rasch Models
Abstract	Scores estimated from multidimensional item response theory (IRT) models are not necessarily comparable across dimensions. In this article, the concept of aligned dimensions is formalized in the context of Rasch models, and two methods are described—delta dimensional alignment (DDA) and logistic regression alignment (LRA)—to transform estimated item parameters so that dimensions are aligned. Both the DDA and LRA methods are applied to real and simulated data, and it is demonstrated that both methods are broadly effective for achieving aligned scales. The routine use of scale alignment methods is recommended prior to comparing scores across dimensions.
Citation	Feuerstahler, L., & Wilson, M. (2019). Scale alignment in between-item multidimensional Rasch models. *Journal of Educational Measurement*, *56*(2), 280–301. https://doi.org/10.1111/jedm.12209
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12209

Title	Standard Errors of IRT Parameter Scale Transformation Coefficients: Comparison of Bootstrap Method, Delta Method, and Multiple Imputation Method
Abstract	The present study evaluated the multiple imputation method, a procedure that is similar to the one suggested by Li and Lissitz (2004), and compared the performance of this method with that of the bootstrap method and the delta method in obtaining the standard errors for the estimates of the parameter scale transformation coefficients in item response theory (IRT) equating in the context of the common-item nonequivalent groups design. Two different estimation procedures for the variance-covariance matrix of the IRT item parameter estimates, which were used in both the delta method and the multiple imputation method, were considered: empirical cross-product (XPD) and supplemented expectation maximization (SEM). The results of the analyses with simulated and real data indicate that the multiple imputation method generally produced very similar results to the bootstrap method and the delta method in most of the conditions. The differences between the estimated standard errors obtained by the methods using the XPD matrices and the SEM matrices were very small when the sample size was reasonably large. When the sample size was small, the methods using the XPD matrices appeared to yield slight upward bias for the standard errors of the IRT parameter scale transformation coefficients.
Citation	Zhang, Z., & Zhao, M. (2019). Standard errors of IRT parameter scale transformation coefficients: Comparison of bootstrap method, delta method, and multiple imputation method. *Journal of Educational Measurement*, *56*(2), 302–330. https://doi.org/10.1111/jedm.12210
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12210

Title	Effectiveness of Equating at the Passing Score for Exams With Small Sample Sizes
Abstract	This article explores the amount of equating error at a passing score when equating scores from exams with small samples sizes. This article focuses on equating using classical test theory methods of Tucker linear, Levine linear, frequency estimation, and chained equipercentile equating. Both simulation and real data studies were used in the investigation. The results of the study supported past findings that as the sample sizes increase, the amount of bias in the equating at the passing score decreases. The research also highlights the importance for practitioners to understand the data, to have an informed expectation of the results, and to have a documented rationale for an acceptable amount of equating error.
Citation	Wolkowitz, A. A., & Wright, K. D. (2019). Effectiveness of equating at the passing score for exams with small sample sizes. *Journal of Educational Measurement*, *56*(2), 361–390. https://doi.org/10.1111/jedm.12212
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12212

Title	Item Calibration Methods With Multiple Subscale Multistage Testing
Abstract	Many large-scale educational surveys have moved from linear form design to multistage testing (MST) design. One advantage of MST is that it can provide more accurate latent trait (θ) estimates using fewer items than required by linear tests. However, MST generates incomplete response data by design; hence, questions remain as to how to calibrate items using the incomplete data from MST design. Further complication arises when there are multiple correlated subscales per test, and when items from different subscales need to be calibrated according to their respective score reporting metric. The current calibration-per-subscale method produced biased item parameters, and there is no available method for resolving the challenge. Deriving from the missing data principle, we showed when calibrating all items together the Rubin's ignorability assumption is satisfied such that the traditional single-group calibration is sufficient. When calibrating items per subscale, we proposed a simple modification to the current calibration-per-subscale method that helps reinstate the missing-at-random assumption and therefore corrects for the estimation bias that is otherwise existent. Three mainstream calibration methods are discussed in the context of MST, they are the marginal maximum likelihood estimation, the expectation maximization method, and the fixed parameter calibration. An extensive simulation study is conducted and a real data example from NAEP is analyzed to provide convincing empirical evidence.
Citation	Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*, *57*(1), 3–28. https://doi.org/10.1111/jedm.12241
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12241

Title	Two IRT Fixed Parameter Calibration Methods for the Bifactor Model
Abstract	New items are often evaluated prior to their operational use to obtain item response theory (IRT) item parameter estimates for quality control purposes. Fixed parameter calibration is one linking method that is widely used to estimate parameters for new items and place them on the desired scale. This article provides detailed descriptions of two fixed parameter calibration methods for the bifactor model and compares their relative performance through simulation. The two methods, which were natural generalizations of their counterparts in the unidimensional context, are the one prior weights updating and multiple expectation-maximization (EM) cycles (OWU-MEM) and multiple prior weights updating and multiple EM cycles (MWU-MEM) methods. In addition, for comparison purposes, the separate calibration method with Haebara linking was included in the simulation. In general, the MWU-MEM method recovered item parameters well for both equivalent and nonequivalent groups, whereas the OWU-MEM method worked well only for equivalent groups. With a few exceptions, the MWU-MEM and Haebara methods showed comparable item parameter recovery.
Citation	Kim, K. Y. (2020). Two IRT fixed parameter calibration methods for the bifactor model. *Journal of Educational Measurement*, *57*(1), 29–50. https://doi.org/10.1111/jedm.12230
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12230

Title	A New Statistic to Assess Fitness of Cubic-Spline Postsmoothing
Abstract	In equating, smoothing techniques are frequently used to diminish sampling error. There are typically two types of smoothing: presmoothing and postsmoothing. For polynomial log-linear presmoothing, an optimum smoothing degree can be determined statistically based on the Akaike information criterion or Chi-square difference criterion. For cubic-spline postsmoothing, visual inspection has been an important tool in choosing such optimum degrees in operational settings. This study introduces a new statistic for assessing the fitness of the cubic-spline postsmoothing method, which accommodates three conditions: (1) one standard error band, (2) deviation from unsmoothed equivalents, and (3) smoothness. A principal advantage of the new statistic proposed in this study is that an optimum degree of smoothing can be selected automatically by giving consistent amount of attention to deviation and smoothness across multiple equatings, whereas visual inspection may not be consistent.
Citation	Kim, H. J., Brennan, R. L., & Lee, W.-C. (2020), A new statistic to assess fitness of cubic-spline postsmoothing. *Journal of Educational Measurement*, *57*(1), 124–144. https://doi.org/10.1111/jedm.12244
Link	https://doi.org/10.1111/jedm.12244

Title	A New Statistic for Selecting the Smoothing Parameter for Polynomial Loglinear Equating Under the Random Groups Design
Abstract	Smoothing is designed to yield smoother equating results that can reduce random equating error without introducing very much systematic error. The main objective of this study is to propose a new statistic and to compare its performance to the performance of the Akaike information criterion and likelihood ratio chi-square difference statistics in selecting the smoothing parameter for polynomial loglinear equating under the random groups design. These model selection statistics were compared for four sample sizes (500, 1,000, 2,000, and 3,000) and eight simulated equating conditions, including both conditions where equating is not needed and conditions where equating is needed. The results suggest that all model selection statistics tend to improve the equating accuracy by reducing the total equating error. The new statistic tended to have less overall error than the other two methods.
Citation	Liu, C., & Kolen, M. J. (2020). A new statistic for selecting the smoothing parameter for polynomial loglinear equating under the random groups design. *Journal of Educational Measurement*, *57*(3), 458–479. https://doi.org/10.1111/jedm.12257
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12257

Title	Linking via Pseudo-Equivalent Group Design: Methodological Considerations and an Application to the PISA and PIAAC Assessments
Abstract	This article presents the pseudo-equivalent group approach and discusses how it can enhance the quality of linking in the presence of nonequivalent groups. The pseudo-equivalent group approach allows to achieve pseudo-equivalence using propensity score reweighting techniques. We use it to perform linking to establish scale concordance between two assessments. The article presents Monte-Carlo simulations and a real data application based on data from the Survey of Adult Skills (PIAAC) and the Programme for International Student Assessment (PISA). Monte-Carlo simulations suggest that the pseudo-equivalent group design is particularly useful whenever there is a large overlap across the two groups with respect to balancing variables and when the correlation between such variables and ability is medium or high. The example based on PISA and PIAAC data indicates that the approach can provide reasonable accurate linking that can be used for group-level comparisons.
Citation	Pokropek, A., & Borgonovi, F. (2020). Linking via pseudo-equivalent group design: Methodological considerations and an application to the PISA and PIAAC assessments. *Journal of Educational Measurement*, *57*(4), 527–546. https://doi.org/10.1111/jedm.12261
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12261

Title	Using a Projection IRT Method for Vertical Scaling When Construct Shift Is Present
Abstract	In vertical scaling, results of tests from several different grade levels are placed on a common scale. Most vertical scaling methodologies rely heavily on the assumption that the construct being measured is unidimensional. In many testing situations, however, such an assumption could be problematic. For instance, the construct measured at one grade level may differ from that measured in another grade (e.g., construct shift). On the other hand, dimensions that involve low-level skills are usually mastered by almost all students as they progress to higher grades. These types of changes in the multidimensional structure, within and across grades, create challenges for developing a vertical scale. In this article, we propose the use of projective IRT (PIRT) as a potential solution to the problem. Assuming that a test measures a primary dimension of substantive interest as well as some peripheral dimensions, the idea underlying PIRT is to integrate out the secondary dimensions such that the model provides both item parameters and ability estimates for the primary dimension. A simulation study was conducted to evaluate the effectiveness of the PIRT as a method for vertical scaling. An example using empirical data from a measure of foundational reading skills is also presented.
Citation	Strachan, T., Cho, U. H., Kim, K. Y., Willse, J. T., Chen, S. H., Ip, E. H., ... Weeks, J. P. (2021). Using a projection IRT method for vertical scaling when construct shift is present. *Journal of Educational Measurement*, *58*(2), 211–235. https://doi.org/10.1111/jedm.12278
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12278

Title	Constructing a Robust Score Scale from IRT Scores with Informed Boundaries
Abstract	In operational testing, item response theory (IRT) models for dichotomous responses are popular for measuring a single latent construct θ, such as cognitive ability in a content domain. Estimates of θ, also called IRT scores or θˆ, can be computed using estimators based on the likelihood function, such as maximum likelihood (ML), weighted likelihood (WL), maximum a posteriori (MAP), and expected a posteriori (EAP). Although the parameter space of θ is theoretically unrestricted, the range of finite θˆ is constrained by the estimator and test form properties, which is important to consider but often overlooked when developing a score scale for reporting purposes. Irrespective of the estimator or test forms at hand, a common practice is to fix arbitrary points symmetric about zero (e.g., −4 and 4) as anchors for deriving a score transformation, possibly resulting in unintended gaps or truncations at the extremes. Therefore, a systematic framework is proposed for using IRT scores to construct a robust score scale with informed boundaries that are logical and consistent across test forms.
Citation	Choe, E. M., & Han, K. C. T. (2022). Constructing a robust score scale from IRT scores with informed boundaries. *Journal of Educational Measurement*, *59*(1), 4–21. https://doi.org/10.1111/jedm.12307
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12307

Title	Assessing the Impact of Equating Error on Group Means and Group Mean Differences
Abstract	Equating error is usually small relative to the magnitude of measurement error, but it could be one of the major sources of error contributing to mean scores of large groups in educational measurement, such as the year-to-year state mean score fluctuations. Though testing programs may routinely calculate the standard error of equating (SEE), the commonly used summary statistics of SEE are not direct quantifications of the impact of equating error on group means. This article proposed summary statistics that directly quantify the impact of equating error on group means or group mean differences and provided empirical and analytical methods to estimate these statistics. Examples based on empirical data were used to illustrate practical applications of these statistics.
Citation	Li, D. (2022). Assessing the impact of equating error on group means and group mean differences. *Journal of Educational Measurement*, *59*(1), 62–79. https://doi.org/10.1111/jedm.12311
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12311

Title	Score Comparability between Online Proctored and In-Person Credentialing Exams
Abstract	This article studied two methods to detect mode effects in two credentialing exams. In Study 1, we used a "modal scale comparison approach," where the same pool of items was calibrated separately, without transformation, within two TC cohorts (TC1 and TC2) and one OP cohort (OP1) matched on their pool-based scale score distributions. The calibrations from all three groups were used to score the TC2 cohort, designated the validation sample. The TC1 item parameters and TC1-based thetas and pass rates were more like the native TC2 values than the OP1-based values, indicating mode effects, but the score and pass/fail decision differences were small. In Study 2, we used a "cross-modal repeater approach" in which test takers who failed their first attempt in one modality took the test again in either the same or different modality. The two pairs of repeater groups (TC → TC: TC → OP, and OP → OP: OP → TC) were matched exactly on their first attempt scores. Results showed increased pass rate and greater score variability in all conditions involving OP, with mode effects noticeable in both the TC → OP condition and less-strongly in the OP → TC condition. Limitations of the study and implications for exam developers were discussed.
Citation	Jones, P., Tong, Y., Liu, J., Borglum, J., & Primoli, V. (2022). Score comparability between online proctored and in-person credentialing exams. *Journal of Educational Measurement*, *59*(2), 180–207. https://doi.org/10.1111/jedm.12320
Link	https://onlinelibrary.wiley.com/doi/10.1111/jedm.12320

Psychometrika: 2010–2022

Title	Accumulative Equating Error after a Chain of Linear Equatings
Abstract	After many equatings have been conducted in a testing program, equating errors can accumulate to a degree that is not negligible compared to the standard error of measurement. In this paper, the author investigates the asymptotic accumulative standard error of equating (ASEE) for linear equating methods, including chained linear, Tucker, and Levine, under the nonequivalent groups with anchor test (NEAT) design. A recursive formula for the ASEE is provided for a series of equatings that makes use of only historical summary statistics. This formula can serve as a new tool to measure the magnitude of equating errors that have accumulated over a series of equatings, and to help monitor and design testing programs.
Citation	Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika*, *75*, 438–453. https://doi.org/10.1007/s11336-010-9160-x
Link	https://link.springer.com/article/10.1007/s11336-010-9160-x

Title	New Equating Methods and Their Relationships with Levine Observed Score Linear Equating Under the Kernel Equating Framework
Abstract	In this paper, we develop a new curvilinear equating for the nonequivalent groups with anchor test (NEAT) design under the assumption of the classical test theory model, that we name curvilinear Levine observed score equating. In fact, by applying both the kernel equating framework and the mean preserving linear transformation of post-stratification equating, we obtain a family of observed score equipercentile equating functions, which also includes the classical Levine observed score linear equating and the Tucker linear equating as special cases.
Citation	Chen, H., & Holland, P. (2010). New equating methods and their relationships with Levine observed score linear equating under the kernel equating framework. *Psychometrika*, *75*, 542–557. https://doi.org/10.1007/s11336-010-9171-7
Link	https://link.springer.com/article/10.1007/s11336-010-9171-7

Title	IRT Test Equating in Complex Linkage Plans
Abstract	Linkage plans can be rather complex, including many forms, several links, and the connection of forms through different paths. This article studies item response theory equating methods for complex linkage plans when the common-item nonequivalent group design is used. An efficient way to average equating coefficients that link the same two forms through different paths will be presented and the asymptotic standard errors of indirect and average equating coefficients are derived. The methodology is illustrated using simulations studies and a real data example.
Citation	Battauz, M. (2013). IRT test equating in complex linkage plans. *Psychometrika*, *78*, 464–480. https://doi.org/10.1007/s11336-012-9316-y
Link	https://link.springer.com/article/10.1007/s11336-012-9316-y

Title	Harmonic Regression and Scale Stability
Abstract	Monitoring a very frequently administered educational test with a relatively short history of stable operation imposes a number of challenges. Test scores usually vary by season, and the frequency of administration of such educational tests is also seasonal. Although it is important to react to unreasonable changes in the distributions of test scores in a timely fashion, it is not a simple matter to ascertain what sort of distribution is really unusual. Many commonly used approaches for seasonal adjustment are designed for time series with evenly spaced observations that span many years and, therefore, are inappropriate for data from such educational tests. Harmonic regression, a seasonal-adjustment method, can be useful in monitoring scale stability when the number of years available is limited and when the observations are unevenly spaced. Additional forms of adjustments can be included to account for variability in test scores due to different sources of population variations. To illustrate, real data are considered from an international language assessment.
Citation	Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*, 815–829. https://doi.org/10.1007/s11336-013-9337-1
Link	https://link.springer.com/article/10.1007/s11336-013-9337-1

Title	Lord–Wingersky Algorithm Version 2.0 for Hierarchical Item Factor Models with Applications in Test Scoring, Scale Alignment, and Model Fit Testing
Abstract	Lord and Wingersky's (Appl Psychol Meas 8:453–461, 1984) recursive algorithm for creating summed score based likelihoods and posteriors has a proven track record in unidimensional item response theory (IRT) applications. Extending the recursive algorithm to handle multidimensionality is relatively simple, especially with fixed quadrature because the recursions can be defined on a grid formed by direct products of quadrature points. However, the increase in computational burden remains exponential in the number of dimensions, making the implementation of the recursive algorithm cumbersome for truly high-dimensional models. In this paper, a dimension reduction method that is specific to the Lord–Wingersky recursions is developed. This method can take advantage of the restrictions implied by hierarchical item factor models, e.g., the bifactor model, the testlet model, or the two-tier model, such that a version of the Lord–Wingersky recursive algorithm can operate on a dramatically reduced set of quadrature points. For instance, in a bifactor model, the dimension of integration is always equal to 2, regardless of the number of factors. The new algorithm not only provides an effective mechanism to produce summed score to IRT scaled score translation tables properly adjusted for residual dependence, but leads to new applications in test scoring, linking, and model fit checking as well. Simulated and empirical examples are used to illustrate the new applications.
Citation	Cai, L. (2015). Lord–Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, *80*, 535–559. https://doi.org/10.1007/s11336-014-9411-3
Link	https://link.springer.com/article/10.1007/s11336-014-9411-3

Title	Linking Item Response Model Parameters
Abstract	With a few exceptions, the problem of linking item response model parameters from different item calibrations has been conceptualized as an instance of the problem of test equating scores on different test forms. This paper argues, however, that the use of item response models does not require any test score equating. Instead, it involves the necessity of parameter linking due to a fundamental problem inherent in the formal nature of these models—their general lack of identifiability. More specifically, item response model parameters need to be linked to adjust for the different effects of the identifiability restrictions used in separate item calibrations. Our main theorems characterize the formal nature of these linking functions for monotone, continuous response models, derive their specific shapes for different parameterizations of the 3PL model, and show how to identify them from the parameter values of the common items or persons in different linking designs.
Citation	van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika*, *81*, 650–673. https://doi.org/10.1007/s11336-015-9469-6
Link	https://link.springer.com/article/10.1007/s11336-015-9469-6

Title	Multiple Equating of Separate IRT Calibrations
Abstract	When test forms are calibrated separately, item response theory parameters are not comparable because they are expressed on different measurement scales. The equating process includes the conversion of item parameter estimates on a common scale and the determination of comparable test scores. Various statistical methods have been proposed to perform equating between two test forms. This paper provides a generalization to multiple test forms of the mean-geometric mean, the mean-mean, the Haebara, and the Stocking–Lord methods. The proposed methods estimate simultaneously the equating coefficients that permit the scale transformation of the parameters of all forms to the scale of the base form. Asymptotic standard errors of the equating coefficients are derived. A simulation study is presented to illustrate the performance of the methods.
Citation	Battauz, M. (2017). Multiple equating of separate IRT calibrations. *Psychometrika*, *82*, 610–636. https://doi.org/10.1007/s11336-016-9517-x
Link	https://link.springer.com/article/10.1007/s11336-016-9517-x

Title	Latent Feature Extraction for Process Data via Multidimensional Scaling
Abstract	Computer-based interactive items have become prevalent in recent educational assessments. In such items, detailed human–computer interactive process, known as response process, is recorded in a log file. The recorded response processes provide great opportunities to understand individuals' problem solving processes. However, difficulties exist in analyzing these data as they are high-dimensional sequences in a nonstandard format. This paper aims at extracting useful information from response processes. In particular, we consider an exploratory analysis that extracts latent variables from process data through a multidimensional scaling framework. A dissimilarity measure is described to quantify the discrepancy between two response processes. The proposed method is applied to both simulated data and real process data from 14 PSTRE items in PIAAC 2012. A prediction procedure is used to examine the information contained in the extracted latent variables. We find that the extracted latent variables preserve a substantial amount of information in the process and have reasonable interpretability. We also empirically prove that process data contains more information than classic binary item responses in terms of out-of-sample prediction of many variables.
Citation	Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, *85*, 378–397. https://doi.org/10.1007/s11336-020-09708-3
Link	https://link.springer.com/article/10.1007/s11336-020-09708-3

Title	On the Behaviour of K-Means Clustering of a Dissimilarity Matrix by Means of Full Multidimensional Scaling
Abstract	In this article, we analyse the usefulness of multidimensional scaling in relation to performing K-means clustering on a dissimilarity matrix, when the dimensionality of the objects is unknown. In this situation, traditional algorithms cannot be used, and so K-means clustering procedures are being performed directly on the basis of the observed dissimilarity matrix. Furthermore, the application of criteria originally formulated for two-mode data sets to determine the number of clusters depends on their possible reformulation in a one-mode situation. The linear invariance property in K-means clustering for squared dissimilarities, together with the use of multidimensional scaling, is investigated to determine the cluster membership of the observations and to address the problem of selecting the number of clusters in K-means for a dissimilarity matrix. In particular, we analyse the performance of K-means clustering on the full dimensional scaling configuration and on the equivalently partitioned configuration related to a suitable translation of the squared dissimilarities. A Monte Carlo experiment is conducted in which the methodology examined is compared with the results obtained by procedures directly applicable to a dissimilarity matrix.
Citation	Vera, J. F., & Macías, R. (2021). On the behaviour of k-means clustering of a dissimilarity matrix by means of full multidimensional scaling. *Psychometrika*, *86*, 489–513. https://doi.org/10.1007/s11336-021-09757-2
Link	https://link.springer.com/article/10.1007/s11336-021-09757-2

Title	Matching IRT Models to Patient-Reported Outcomes Constructs: The Graded Response and Log-Logistic Models for Scaling Depression
Abstract	Item response theory (IRT) model applications extend well beyond cognitive ability testing, and various patient-reported outcomes (PRO) measures are among the more prominent examples. PRO (and like) constructs differ from cognitive ability constructs in many ways, and these differences have model fitting implications. With a few notable exceptions, however, most IRT applications to PRO constructs rely on traditional IRT models, such as the graded response model. We review some notable differences between cognitive and PRO constructs and how these differences can present challenges for traditional IRT model applications. We then apply two models (the traditional graded response model and an alternative log-logistic model) to depression measure data drawn from the Patient-Reported Outcomes Measurement Information System project. We do not claim that one model is "a better fit" or more "valid" than the other; rather, we show that the log-logistic model may be more consistent with the construct of depression as a unipolar phenomenon. Clearly, the graded response and log-logistic models can lead to different conclusions about the psychometrics of an instrument and the scaling of individual differences. We underscore, too, that, in general, explorations of which model may be more appropriate cannot be decided only by fit index comparisons; these decisions may require the integration of psychometrics with theory and research findings on the construct of interest.
Citation	Reise, S. P., Du, H., Wong, E. F. et al. (2021). Matching IRT models to patient-reported outcomes constructs: The graded response and log-logistic models for scaling depression. *Psychometrika*, *86*, 800–824. https://doi.org/10.1007/s11336-021-09802-0
Link	https://link.springer.com/article/10.1007/s11336-021-09802-0

Title	Linking Scores with Patient-Reported Health Outcome Instruments: A Validation Study and Comparison of Three Linking Methods
Abstract	The psychometric process used to establish a relationship between the scores of two (or more) instruments is generically referred to as linking. When two instruments with the same content and statistical test specifications are linked, these instruments are said to be equated. Linking and equating procedures have long been used for practical benefit in educational testing. In recent years, health outcome researchers have increasingly applied linking techniques to patient-reported outcome (PRO) data. However, these applications have some noteworthy purposes and associated methodological questions. Purposes for linking health outcomes include the harmonization of data across studies or settings (enabling increased power in hypothesis testing), the aggregation of summed score data by means of score crosswalk tables, and score conversion in clinical settings where new instruments are introduced, but an interpretable connection to historical data is needed. When two PRO instruments are linked, assumptions for equating are typically not met and the extent to which those assumptions are violated becomes a decision point around how (and whether) to proceed with linking. We demonstrate multiple linking procedures—equipercentile, unidimensional IRT calibration, and calibrated projection—with the Patient-Reported Outcomes Measurement Information System Depression bank and the Patient Health Questionnaire-9. We validate this link across two samples and simulate different instrument correlation levels to provide guidance around which linking method is preferred. Finally, we discuss some remaining issues and directions for psychometric research in linking PRO instruments.
Citation	Schalet, B. D., Lim, S., Cella, D. et al. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika*, *86*, 717–746. https://doi.org/10.1007/s11336-021-09776-z
Link	https://link.springer.com/article/10.1007/s11336-021-09776-z

Selected Research Reports: ETS

Title	An Empirical Comparison of Methods for Equating With Randomly Equivalent Groups of 50 to 400 Test Takers
Abstract	A series of resampling studies investigated the accuracy of equating by four different methods in a random groups equating design with samples of 400, 200, 100, and 50 test takers taking each form. Six pairs of forms were constructed. Each pair was constructed by assigning items from an existing test taken by 9,000 or more test takers. The criterion equating was the direct equipercentile equating in the full group. Accuracy was described in terms of the root-mean-squared deviation (over 1,000 replications) of the sample equatings from the criterion equating. The equating methods investigated were equipercentile equating of smoothed distributions, linear equating, mean equating, and circle-arc equating; they were compared with each other and with the identity. Circle-arc equating produced the most accurate results for all sample sizes investigated, particularly in the upper half of the score distribution.
Citation	Livingston, S. A., & Kim, S. (2010). *An empirical comparison of methods for equating with randomly equivalent groups of 50 to 400 test takers* (Research Report No. RR-10-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02212.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02212.x

Title	Chained Versus Post-Stratification Equating in a Linear Context: An Evaluation Using Empirical Data
Abstract	This study used real data to construct testing conditions for comparing results of chained linear, Tucker, and Levine-observed score equatings. The comparisons were made under conditions where the new- and old-form samples were similar in ability and when they differed in ability. The length of the anchor test was also varied to enable examination of its effect on the three different equating methods. Two tests were used in the study, and the three equating methods were compared to a criterion equating to obtain estimates of random equating error, bias, and root mean squared error (RMSE). Results showed that for most of the conditions studied, chained linear score equating produced fairly good equating results in terms of low bias and RMSE. In some conditions, Levine-observed score equating also produced low bias and RMSE. Although the Tucker method always produced the lowest random equating error, it produced a larger bias and RMSE than either of the other equating methods. Based on these results, it is recommended that either chained linear or Levine score equating be used when new- and old-form samples differ in ability and/or when the anchor-to-total correlation is not very high.
Citation	Puhan, G. (2010). *Chained versus post-stratification equating in a linear context: An evaluation using empirical data* (Research Report No. RR-10-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02213.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02213.x

Title	Aligning Scales of Certification Tests
Abstract	Scores are the most visible and widely used products of a testing program. The choice of score scale has implications for test specifications, equating, and test reliability and validity, as well as for test interpretation. At the same time, the score scale should be viewed as infrastructure likely to require repair at some point. In this report we examine the issue of scale fit—how well the scale fits the intended uses of its scores—for certification tests. Two examples of scale fit are considered: one in which the test has a single threshold that separates the candidate population into pass-fail groups, and one in which the test is required to support a restricted range of multiple thresholds.
Citation	Dorans, N. J., Liang, L., & Puhan, G. (2010). *Aligning scales of certification tests* (Research Report No. RR-10-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02214.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02214.x

Title	Linking Errors in Trend Estimation in Large-Scale Surveys: A Case Study
Abstract	One of the major objectives of large-scale educational surveys is reporting trends in academic achievement. For this purpose, a substantial number of items are carried from one assessment cycle to the next. The linking process that places academic abilities measured in different assessments on a common scale is usually based on a concurrent calibration of adjacent assessments using item response theory (IRT) models. It can be conjectured that the selection of common items has a direct effect on the estimation error of academic abilities due to item misfit, small changes in the common items, position effect, and other sources of construct-irrelevant changes between measurement occasions. Hence, the error due to the common-item sampling could be a major source of error for the ability estimates. In operational analyses, generally two sources of error are accounted for in variance estimation: student sampling error and measurement error. A double jackknifing procedure is proposed to include a third source of the estimation error, the error due to common-item sampling. Three different versions of the double jackknifing were implemented and compared. The data used in this study were item responses from Grade 4 students who took the NAEP 2004 and 2008 math long-term trend (LTT) assessments. These student samples used in this study are representative samples of Grade 4 student population in 2004 and in 2008 across the US. The results showed that these three double jackknifing approaches resulted in similar standard error estimates that were slightly higher than the estimates from the traditional approach, regardless of whether an item sampling scheme was used or items were dropped at random.
Citation	Xu, X., & von Davier, M. (2010). *Linking errors in trend estimation in large-scale surveys: A case study* (Research Report No. RR-10-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02217.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02217.x

Title	Single- Versus Double-Scoring of Trend Responses in Trend Score Equating With Constructed-Response Tests
Abstract	This study examines the differences in equating outcomes between two trend score equating designs resulting from two different scoring strategies for trend scoring when operational constructed-response (CR) items are double-scored—the single group (SG) design, where each trend CR item is double-scored, and the nonequivalent groups with anchor test (NEAT) design, where each trend CR item is single-scored during trend score equating—for varying sample sizes (n =150, 200, 250, 300, 400). Overall results suggest larger equating errors with smaller sample sizes, though errors were small regardless of sample size. The NEAT design performed about as well as the SG design with respect to conditional and summative standard errors of equating, though it did tend to produce larger bias and root mean-squared differences (RMSDs). When accounting for the total number of trend scores required to do analyses, the NEAT design performed as well or better than the SG design (e.g., when the NEAT n =150 and the SG n = 300). This result might be partially attributable to a larger operational sample size (n = 792) and a good correlation between anchor and total score for the trend sample (r = 0.73). These results suggest that under these testing conditions, the NEAT design performed about as well as the SG design, but further research is required to assess the generalizability of the results.
Citation	Tan, X., Ricker, K. L., & Puhan, G. (2010). *Single- versus double-scoring of trend responses in trend score equating with constructed-response tests* (Research Report No. RR-10-12). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02219.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02219.x

Title	Examining Two Strategies to Link Mixed-Format Tests Using Multiple-Choice Anchors
Abstract	This study examined the use of an all multiple-choice (MC) anchor for linking mixed format tests containing both MC and constructed-response (CR) items, in a nonequivalent groups design. An MC-only anchor could effectively link two such test forms if either (a) the MC and CR portions of the test measured the same construct, so that the MC anchor adequately represented the entire test, or (b) the relationship between the MC portion and the total test remained constant across the new and reference linking groups. The study also evaluated whether linking mixed-format tests through MC-only anchors would be more effective than a two-stage strategy in which MC portions were equated through MC anchors and then composite scores were scaled to the MC scores. Anchor linking and two-stage linking yielded identical (or nearly so) results for both linear and nonlinear chained linking methods. With post-stratification linking methods the two-stage strategy resulted in smaller bias. The paper discusses some advantages of both approaches.
Citation	Walker, M. E., & Kim, S. (2010). *Examining two strategies to link mixed-format tests using multiple-choice anchors* (Research Report No. RR-10-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02225.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02225.x

Title	Limits on the Accuracy of Linking
Abstract	Sampling errors limit the accuracy with which forms can be linked. Limitations on accuracy are especially important in testing programs in which a very large number of forms are employed. Standard inequalities in mathematical statistics may be used to establish lower bounds on the achievable inking accuracy. To illustrate results, a variety of equating problems are considered.
Citation	Haberman, S. J. (2010). *Limits on the accuracy of linking* (Research Report No. RR-10-22). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02229.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02229.x

Title	The Use of Two Anchors in Nonequivalent Groups With Anchor Test (NEAT) Equating
Abstract	In the equating literature, a recurring concern is that equating functions that utilize a single anchor to account for examinee groups' nonequivalence are biased when the groups are extremely different and/or when the anchor only weakly measures what the tests measure. Several proposals have been made to address this equating bias by incorporating more than one anchor into nonequivalent groups with anchor test (NEAT) equating functions. These proposals have not been extensively considered or comparatively evaluated. This study evaluates three methods for incorporating more than one anchor into NEAT equating functions, including poststratification, imputation, and propensity score matching. The three methods are studied and compared in two examples. The implications for using the three equating approaches in practice and for developing alternative strategies to incorporate two anchors are discussed.
Citation	Moses, T., Deng, W., & Zhang, Y.-L. (2010). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating* (Research Report No. RR-10-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02230.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02230.x

Title	Research on Standard Errors of Equating Differences
Abstract	In this paper, the standard error of equating difference (SEED) is described in terms of originally proposed kernel equating functions (von Davier, Holland & Thayer, 2004) and extended to incorporate traditional linear and equipercentile functions. These derivations expand on prior developments of SEEDs and standard errors of equating and provide additional insight about the relationships of kernel and traditional equating functions. Simulations are used to evaluate the SEEDs' accuracies.
Citation	Moses, T., & Zhang, W. (2010). *Research on standard errors of equating differences* (Research Report No. RR-10-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02232.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02232.x

Title	Principles and Practices of Test Score Equating
Abstract	Score equating is essential for any testing program that continually produces new editions of a test and for which the expectation is that scores from these editions have the same meaning over time. Particularly in testing programs that help make high-stakes decisions, it is extremely important that test equating be done carefully and accurately. An error in the equating function or score conversion can affect the scores for all examinees, which is both a fairness and a validity concern. Because the reported score is so visible, the credibility of a testing organization hinges on activities associated with producing, equating, and reporting scores. This paper addresses the practical implications of score equating by describing aspects of equating and best practices associated with the equating process.
Citation	Dorans, N. J., Moses, T., & Eignor, D. R. (2010). *Research on standard errors of equating differences* (Research Report No. RR-10-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02236.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2010.tb02236.x

Title	Equating of Subscores and Weighted Averages Under the NEAT Design
Abstract	Recently, the literature has seen increasing interest in subscores for their potential diagnostic values; for example, one study suggested the report of weighted averages of a subscore and the total score, whereas others showed, for various operational and simulated data sets, that weighted averages, as compared to subscores, lead to more accurate diagnostic information. To report weighted averages, the averages should be comparable across different test forms; that is, the averages should be equated. This report discusses how to equate weighted averages. Results from operational and simulated data sets demonstrate the small error found when equating weighted averages.
Citation	Sinharay, S., & Haberman, S. (2011). *Equating of subscores and weighted averages under the NEAT design* (Research Report No. RR-11-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02237.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02237.x

Title	Statistical Procedures to Evaluate Quality of Scale Anchoring
Abstract	Providing information to test takers and test score users about the abilities of test taker sat different score levels has been a persistent problem in educational and psychological measurement (Carroll, 1993). Scale anchoring (Beaton & Allen, 1992), a technique that describes what students at different points on a score scale know and can do, is a tool to provide such information. Scale anchoring for a test involves substantial amount of work, both by the statistical analysts and test developers involved with the test. In addition, scale anchoring involves considerable use of subjective judgment, so its conclusions may be questionable. This paper describes statistical procedures that can be used to determine if scale anchoring is likely to be successful for a test. If these procedures indicate that scale anchoring is unlikely to be successful, then there is little reason to perform a detailed scale anchoring study. The procedures are applied to several data sets from a teacher licensing test.
Citation	Haberman, S., Sinharay, S., & Lee, Y.-H. (2011). *Statistical procedures to evaluate quality of scale anchoring* (Research Report No. RR-11-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02238.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02238.x

Title	Assessing the Falsifiability of Extreme Linking
Abstract	Extreme linkings are performed in settings in which neither equivalent groups nor anchor material is available to link scores on two assessments. Examples of extreme linkages include links between scores on tests administered in different languages or between scores on tests administered across disability groups. The strength of interpretation attached to a linkage depends on the proper design and execution of a sound data collection plan. The current paper uses a real data set to illustrate how to indirectly assess the quality of linking the scores on two assessments that contain neither equivalent groups nor common anchor material.
Citation	Middleton, K., & Dorans, N. J. (2011). *Assessing the falsifiability of extreme linking* (Research Report No. RR-11-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02240.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02240.x

Title	Equating Subscores Using Total Scaled Scores as an Anchor
Abstract	Because the demand for subscores is ever increasing, this study examined two different approaches for equating subscores: (a) equating a subscore on the new form to the same subscore in the old form using internal common items as the anchor to conduct the equating, and (b) equating a subscore on the new form to the same subscore in the old form using equated total scores as the anchor to conduct the equating. Equated total scores can be used as an anchor to equate the subscores because the total equated scores are comparable across both the new and the old forms. Data from 2 tests (Tests X and Y) were used to conduct the study, and results showed that when the number of internal common items was large (approximately 50% of the total subscore), then using common items to equate the subscores was preferable. However, when the number of common items was small (approximately 25% of the total subscore, which is common practice), then using total scaled scores (TSS) to equate the subscores was preferable. Using raw subscores (not equating) resulted in a considerable amount of bias for both tests.
Citation	Puhan, G., & Liang, L. (2011). *Equating subscores using total scaled scores as an anchor* (Research Report No. RR-11-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02243.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02243.x

Title	Can Smoothing Help When Equating With Unrepresentative Small Samples?
Abstract	The study evaluated the effectiveness of log-linear presmoothing (Holland & Thayer, 1987) on the accuracy of small sample chained equipercentile equatings under two conditions (i.e., using small samples that differed randomly in ability from the target population versus using small samples that were distinctly different from the target population). Results showed that equating with small samples (e.g., $N < 50$) using either raw or smoothed score distributions can result in a substantial amount of random equating error (although smoothing reduced random equating error). Even with samples sizes of 100, the random equating error was quite large (greater than the difference that matters or DTM) for almost all score points. Moreover, when the small samples were unrepresentative of the target population, which is quite likely for small samples, the amount of equating bias (in addition to random equating error) was considerably large for both the raw and smoothed equatings. It was concluded that although presmoothing helped reduce random equating error, it is unlikely to reduce equating bias caused by using an unrepresentative sample. Other alternatives to the small sample equating problem that focus more on improving data collection than on improving existing equating methods are discussed.
Citation	Puhan, G. (2011). *Can smoothing help when equating with unrepresentative small samples?* (Research Report No. RR-11-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02245.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02245.x

Title	Sources of Score Scale Inconsistency
Abstract	For testing programs that administer multiple forms within a year and across years, score equating is used to ensure that scores can be used interchangeably. In an ideal world, samples sizes are large and representative of populations that hardly change over time, and very reliable alternate test forms are built with nearly identical psychometric properties. Under these conditions, most equating methods produce score conversions close to the identity function. Unfortunately, equating is sometimes performed on small non-representative samples with variable distributions of ability, and administered tests are built to vague specifications. Here, different equating methods produce different results because they are based on different assumptions. In the nearly ideal case, there are smaller deviations from the identity function because great effort is taken to control variation. Even when equating is conducted under these desirable conditions, the random variation in form-to-form equating, when concatenated over time, can produce substantial shifts in score conversions, that is, scale drift. In this paper, we make distinctions among different sources of variation that may contribute to score-scale inconsistency, and identify practices that are likely to contribute to it.
Citation	Haberman, S. J., & Dorans, N. J. (2011). *Sources of score scale inconsistency* (Research Report No. RR-11-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02246.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02246.x

Title	Use of Continuous Exponential Families to Link Forms via Anchor Tests
Abstract	Continuous exponential families are applied to linking test forms via an internal anchor. This application combines work on continuous exponential families for single-group designs and work on continuous exponential families for equivalent-group designs. Results are compared to those for kernel and equipercentile equating in the case of chained equating. The conversions produced by all methods are quite similar.
Citation	Haberman, S. J., & Yan, D. (2011). *Use of continuous exponential families to link forms via anchor tests* (Research Report No. RR-11-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02247.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02247.x

Title	Smoothing and Equating Methods Applied to Different Types of Test Score Distributions and Evaluated With Respect to Multiple Equating Criteria
Abstract	In equating research and practice, equating functions that are smooth are typically assumed to be more accurate than equating functions with irregularities. This assumption presumes that population test score distributions are relatively smooth. In this study, two examples were used to reconsider common beliefs about smoothing and equating. The first example involves a relatively smooth population test score distribution and the second example involves a population test score distribution with systematic irregularities. Various smoothing and equating methods (presmoothing, equipercentile, kernel, and postsmoothing) were compared across the two examples with respect to how well the test score distributions were reflected in the equating functions, the smoothness of the equating functions, and the standard errors of equating. The smoothing and equating methods performed more similarly in the first example than in the second example. The results of the second example illustrate that when dealing with systematically irregular test score distributions, smoothing and equating methods can be used in different ways to satisfy different equating criteria.
Citation	Moses, T., & Liu, J. (2011). *Smoothing and equating methods applied to different types of test score distributions and evaluated with respect to multiple equating criteria* (Research Report No. RR-11-20). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02256.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02256.x

Title	The Single Group With Nearly Equivalent Tests (SiGNET) Design for Equating Very Small Volume Multiple-Choice Tests
Abstract	The single group with nearly equivalent tests (SiGNET) design proposed here was developed to address the problem of equating scores on multiple-choice test forms with very small single-administration samples. In this design, the majority of items in each new test form consist of items from the previous form, and the new items that were administered as unscored items in the previous form. Each form is equated using data from examinees who took the previous form. As the equating is a single-group design, with the 2 forms having a large number of overlapping items, the size of the equating sample can be much smaller than in other designs.
Citation	Grant, M. C. (2011). *The single group with nearly equivalent tests (SiGNET) design for equating very small volume multiple-choice tests* (Research Report No. RR-11-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02267.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02267.x

Title	Does Linking Mixed-Format Tests Using a Multiple-Choice Anchor Produce Comparable Results for Male and Female Subgroups?
Abstract	This study examines the use of subpopulation invariance indices to evaluate the appropriateness of using a multiple-choice (MC) item anchor in mixed-format tests, which include both MC and constructed-response (CR) items. Linking functions were derived in the nonequivalent groups with anchor test (NEAT) design using an MC-only anchor set for 4 mixed-format licensure tests. For each of those licensure tests, the linking functions were also derived separately for males and females, and those subpopulation functions were compared to the total group function. The mathematics, social studies, and science tests each produced acceptable differences between each of the subpopulation functions and the total group function within the cut-score region, leading to consistent pass/fail designations for the examinees. The English test, which had a low correlation between MC and CR components (indicative of multidimensionality), produced the largest differences, casting doubt on the effectiveness of the MC-only anchor.
Citation	Kim, S., & Walker, M. E. (2011). *Does linking mixed-format tests using a multiple-choice anchor produce comparable results for male and female subgroups?* (Research Report No. RR-11-44). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02280.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02280.x

Title	Multiple Linking in Equating and Random Scale Drift
Abstract	Maintaining score stability is crucial for an ongoing testing program that administers several tests per year over many years. One way to stall the drift of the score scale is to use an equating design with multiple links. In this study, we use the operational and experimental SAT® data collected from 44 administrations to investigate the effect of accumulated equating error in equating conversions and the effect of the use of multiple links in equating. No equating error is directly observed or calculated in the study. Instead, we focus on the behavior of the equating conversions after a series of equatings under the nonequivalent groups with anchor test design and analyze the effect of equating error on conversions. It is observed that the single-link equating conversions drift further away from the operational ones as more equatings are carried out. Analysis of variance is used to decompose the scale score means and the conversion into two major factors: administration month and year for both single- and multiple-link equating results. Seasonality is seen in the data. In addition, the single-link conversions exhibit a certain instability that is not obvious for the operational data. A statistical random walk model is offered to explain the mechanism of scale drift in equating caused by random equating error.
Citation	Guo, H., Liu, J., Dorans, N., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift* (Research Report No. RR-11-46). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02282.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02282.x

Title	Evaluating Academic Progress Without a Vertical Scale
Abstract	Alternatives to vertical scales are compared for measuring longitudinal academic growth and for producing school-level growth measures. The alternatives examined were empirical cross-grade regression, ordinary least squares and logistic regression, and multilevel models. The student data used for the comparisons were Arabic Grades 4 to 10 in Qatar, and results were examined in the scale score and performance level metrics. It is found that vertical scales and cross-grade regressions can show different results at the individual student level, but at the school level, the different measures of growth were strongly correlated, particularly in the scale score metric. Differences between the methods appear more likely in the performance level metric than the scale score metric and for grade pairs with more extreme performance.
Citation	Yen, W. M., Lall, V. F., & Monfils, L. (2012). *Evaluating academic progress without a vertical scale* (Research Report No. RR-12-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02289.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2012.tb02289.x

Title	Examining the Impact of Drifted Polytomous Anchor Items on Test Characteristic Curve (TCC) Linking and IRT True Score Equating
Abstract	In a common-item (anchor) equating design, the common items should be evaluated for item parameter drift. Drifted items are often removed. For a test that contains mostly dichotomous items and only a small number of polytomous items, removing some drifted polytomous anchor items may result in anchor sets that no longer resemble mini-versions of the new and old test forms. In this study, the impact of drifted polytomous anchor items on the test characteristic curve (TCC) linking and item response theory (IRT) true score equating for a test containing only a small number of polytomous items was investigated. Simulated tests were constructed to mimic a real large-scale test. The magnitude of the item parameter drift, anchor length, number of drifted polytomous items in the anchor set, and the ability distributions of the groups taking the old form and new form were manipulated. Results suggest that anchor length and number of drifted polytomous items had a relatively large impact on the linking and equating results. The accuracy of linking and equating results were affected by the magnitude of item parameter drift. The ability distributions of the groups had little effect on the linking and equating results. In general, excluding drifted polytomous anchor items resulted in an improvement in equating results.
Citation	Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating* (Research Report No. RR-12-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02291.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2012.tb02291.x

Title	Does Preequating Work? An Investigation Into a Preequated Testlet-Based College Placement Exam Using Postadministration Data
Abstract	In this study, we investigated whether preequating results agree with equating results that are based on observed operational data (postequating) for a college placement program. Specifically, we examined the degree to which item response theory (IRT) true score preequating results agreed with those from IRT true score postequating and from observed score equating. Three academic subjects were examined in this study: analyzing and interpreting literature, American government, and college algebra. The findings suggested that differences between equating results from IRT true score preequating and postequating varied from subject to subject. In general, IRT true score postequating agreed with IRT true score preequating for most of the forms for a test subject. Any difference among the equating results can be attributed to the way through which items were pretested, contextual/order effects, or the violation of IRT assumptions.
Citation	Gao, R., He, W., & Ruan, C. (2012). *Does preequating work? An investigation into a preequated testlet-based college placement exam using postadministration data* (Research Report No. RR-12-12). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02294.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2012.tb02294.x

Title	A Note on the Choice of an Anchor Test in Equating
Abstract	Anchor tests play a key role in test score equating. We attempt to find, through theoretical derivations, an anchor test with optimal item characteristics. The correlation between the scores on a total test and on an anchor test is maximized with respect to the item parameters for data satisfying several item response theory models. Results suggest that under these models, the minitest, the traditionally used anchor test, is not optimal with respect to anchor-test-to-total-test correlation; instead, an anchor test with items of medium difficulty, the miditest, seems to be the optimum anchor test. This finding agrees with the empirical findings of Sinharay and colleagues that the miditest mostly has higher anchor-test-to-total-test correlation compared to the minitest and mostly performs as well as the minitest in equating.
Citation	Sinharay, S., Haberman, S., Holland, P., & Lewis, C. (2012). *A note on the choice of an anchor test in equating* (Research Report No. RR-12-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02296.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2012.tb02296.x

Title	The Stability of the Score Scales for the SAT Reasoning Test™ From 2005 to 2010
Abstract	This study examines the stability of the SAT Reasoning Test™ score scales from 2005 to 2010. A 2005 old form (OF) was administered along with a 2010 new form (NF). A new conversion for OF was derived through direct equipercentile equating. A comparison of the newly derived and the original OF conversions showed that Critical Reading and Mathematics score scales have experienced, at most, a moderate upward scale drift (no greater than 5 points on average), and the drift may be explained by an accumulation of random equating errors. The Writing score scale has experienced a significant upward scale drift (11 points on average), which may be caused by sources other than random equating errors.
Citation	Guo, H., Liu, J., Curley, E., & Dorans, N. (2012). *The stability of the score scales for the SAT Reasoning Test™ from 2005 to 2010* (Research Report No. RR-12-15). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02297.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2012.tb02297.x

Title	Exploring Alternative Test Form Linking Designs With Modified Equating Sample Size and Anchor Test Length
Abstract	The purpose of this study was to evaluate the combined effects of reduced equating sample size and shortened anchor test length on item response theory (IRT)-based linking and equating results. Data from two independent operational forms of a large-scale testing program were used to establish the baseline results for evaluating the results from two alternative designs. Under the two alternative designs, two simulated conditions were created from the original data. Under one condition, we reduced the equating sample size (from about 2,000 to about 1,000) per anchor item and shortened the anchor test length (by half) per equating sample. Under the other condition, we reduced the sample size (from about 2,000 to about 1,000) per anchor item only. A complete grouped jackknife replication method was used to estimate the standard errors of the linking and equating procedures from 100 jackknife replicate samples; the complete procedures included IRT calibrations, item parameter scaling, and IRT true score equating. The findings from a comparison of the results from the two simulated conditions and the baseline results showed that neither alternative design had any practical impact on the linking and equating results for either test form.
Citation	Wang, L., Qian, J., & Lee, Y.-H. (2013). *Exploring alternative test form linking designs with modified equating sample size and anchor test length* (Research Report No. RR-13-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02309.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02309.x

Title	Choice of Target Population Weights in Rater Comparability Scoring and Equating
Abstract	The purpose of this study was to demonstrate that the choice of sample weights when defining the target population under poststratification equating can be a critical factor in determining the accuracy of the equating results under a unique equating scenario, known as rater comparability scoring and equating. The nature of data collection under rater comparability scoring is such that it results in a very high correlation between the anchor and total score in the new form but only a moderate correlation in the reference form. I demonstrated, using data collected under a rater comparability scoring situation, that this difference in the anchor-total correlation in the new and reference forms can have a predictable impact on the equating results based on different sample weights (i.e., the equating results are most accurate when the reference form sample is defined as the target population, least accurate when the new form sample is defined as the target population, and somewhere in the middle when the new and reference form samples are equally weighed when defining the target population).
Citation	Puhan, G. (2013). *Choice of target population weights in rater comparability scoring and equating* (Research Report No. RR-13-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02310.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02310.x

Title	A Criterion to Evaluate the Individual Raw-to-Scale Equating Conversions
Abstract	In this study we investigated when an equating conversion line is problematic in terms of gaps and clumps. We suggest using the conditional standard error of measurement (CSEM) to measure the scale scores that are inappropriate in the overall raw-to-scale transformation.
Citation	Guo, H., Puhan, G., & Walker, M. (2013). *A criterion to evaluate the individual raw-to-scale equating conversions* (Research Report No. RR-13-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02312.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02312.x

Title	Poststratification Equating Based on True Anchor Scores and Its Relationship to Levine Observed Score Equating
Abstract	This paper presents a new equating method for the nonequivalent groups with anchor test design: poststratification equating based on true anchor scores. The linear version of this method is shown to be equivalent, under certain conditions, to Levine observed score equating, in the same way that the linear version of poststratification equating is equivalent to Tucker equating. Some issues related to this result are discussed.
Citation	Chen, H., & Livingston, S. A. (2013). *Poststratification equating based on true anchor scores and its relationship to Levine observed score equating* (Research Report No. RR-13-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02318.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02318.x

Title	The Kernel Levine Equipercentile Observed-Score Equating Function
Abstract	In the framework of the observed-score equating methods for the nonequivalent groups with anchor test design, there are 3 fundamentally different ways of using the information provided by the anchor scores to equate the scores of a new form to those of an old form. One method uses the anchor scores as a conditioning variable, such as the Tucker method and poststratification equating. A second way to use the anchor scores is as the middle link in a chain of linking relationships, such as chain linear equating and chain equating. The third way to use the anchor scores is in conjunction with the classical test theory, such as Levine observed-score equating and the newly created hybrid Levine equipercentile equating and poststratification equating based on true anchor scores. The purpose of this paper is to demonstrate that with real data, under certain conditions, hybrid Levine equipercentile equating and poststratification equating based on true anchor scores outperform both poststratification equating and chain equating.
Citation	von Davier, A. A., & Chen, H. (2013). *The kernel Levine equipercentile observed-score equating function* (Research Report No. RR-13-38). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02345.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02345.x

Title	Weighting Test Samples in IRT Linking and Equating: Toward an Improved Sampling Design for Complex Equating
Abstract	Several factors could cause variability in item response theory (IRT) linking and equating procedures, such as the variability across examinee samples and/or test items, seasonality, regional differences, native language diversity, gender, and other demographic variables. Hence, the following question arises: Is it possible to select optimal samples of examinees so that the IRT linking and equating can be more precise at an administration level as well as over a large number of administrations? This is a question of optimal sampling design in linking and equating. To obtain an improved sampling design for invariant linking and equating across testing administrations, we applied weighting techniques to yield a weighted sample distribution that is consistent with the target population distribution. The goal is to obtain a stable Stocking-Lord test characteristic curve (TCC) linking and a true-score equating that is invariant across administrations. To study the weighting effects on linking, we first selected multiple subsamples from a data set. We then compared the linking parameters from subsamples with those from the data and examined whether the linking parameters from the weighted sample yielded smaller mean square errors (MSE) than those from the unweighted subsample. To study the weighting effects on true-score equating, we also compared the distributions of the equated scores. Generally, the findings were that the weighting produced good results.
Citation	Qian, J., Jiang, Y., & von Davier, A. A. (2013). *Weighting test samples in IRT linking and equating: Toward an improved sampling design for complex equating* (Research Report No. RR-13-39). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02346.x
Link	https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02346.x

Title	Enhancing the Equating of Item Difficulty Metrics: Estimation of Reference Distribution
Abstract	Two methods are currently in use at Educational Testing Service (ETS) for equating observed item difficulty statistics. The first method involves the linear equating of item statistics in an observed sample to reference statistics on the same items. The second method, or the item response curve (IRC) method, involves the summation of conditional observed item statistics across the reference population total score frequencies. This article introduces a quick and effective method for obtaining the reference distribution for the transition from the linear equating method to the IRC method without recalculating all the item difficulties. More specifically, a mathematical formula is derived to estimate the score distribution of a reference group that maintains the current item difficulty scale. Future research is needed to compare the performance of the two approaches.
Citation	Ali, U. S., & Walker, M. E. (2014). *Enhancing the equating of item difficulty metrics: Estimation of reference distribution* (Research Report No. RR-14-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12006
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12006

Title	A Comparison of Raw-to-Scale Conversion Consistency Between Single- and Multiple-Linking Using a Nonequivalent Groups Anchor Test Design
Abstract	Maintaining score interchangeability and scale consistency is crucial for any testing programs that administer multiple forms across years. The use of a multiple linking design, which involves equating a new form to multiple old forms and averaging the conversions, has been proposed to control scale drift. However, the use of multiple linking often conflicts with the need for minimizing old item/form exposure and the need for pretesting. This study tried to find a balance point where the needs for equating, item/form exposure and controlling, and pretesting can be satisfied. Three equating scenarios were examined using real data: equating to one old form, equating to two old forms, or equating to three old forms. The finding is that the equating based on one old form produced persistent score drift and also showed increased variability in score means and standard deviations over time. In contrast, equating back to two or three old forms produced much more stable conversions and had less variation. Overall, equating based on multiple linking designs shows the promise of producing more consistent results and preventing scale drift. We recommend that testing programs and practitioners consider the use of multiple linking whenever possible.
Citation	Liu, J., Guo, H., & Dorans, N. (2014). *A comparison of raw-to-scale conversion consistency between single- and multiple-linking using a nonequivalent groups anchor test design* (Research Report No. RR-14-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12014
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12014

Title	Test Score Equating Using Discrete Anchor Items Versus Passage-Based Anchor Items: A Case Study Using SAT® Data
Abstract	The purpose of this study is to investigate the impact of discrete anchor items versus passage-based anchor items on observed score equating using empirical data. This study compares an SAT® critical reading anchor that contains more discrete items proportionally, compared to the total tests to be equated, to another anchor that contains fewer discrete items and more passage-based items proportionally. Both of these anchors were administered in an SAT administration. The impact of anchor type on equating was evaluated with respect to equating bias. The results clearly reveal that the anchor with more discrete items almost always leads to more accurate equating functions than does the anchor with more passage-based items.
Citation	Liu, J., Zu, J., Curley, E., & Carey, J. (2014). *Test score equating using discrete anchor items versus passage-based anchor items* (Research Report No. RR-14-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12015
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12015

Title	Simulate to Understand Models, Not Nature
Abstract	Simulations are widely used. Simulations produce numbers that are deductive demonstrations of what a model says will happen. They produce numerical results that are consistent with the premises of the model used to generate the numbers. These simulated numerical results are not empirical data that address aspects of the world that lies outside the model. In contrast, empirical data are central to the scientific method. When a simulation is substituted for the assessment of hypotheses with real data, a false sense of understanding can ensue and with it a biased perspective on the world. To illustrate the limitations of simulation and their proper role, examples are drawn from simulation studies about score equating.
Citation	Dorans, N. J. (2014). *Simulate to understand models, not nature* (Research Report No. RR-14-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12013
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12013

Title	Effect of Item Response Theory (IRT) Model Selection on Testlet-Based Test Equating
Abstract	The local item independence assumption underlying traditional item response theory (IRT) models is often not met for tests composed of testlets. There are 3 major approaches to addressing this issue: (a) ignore the violation and use a dichotomous IRT model (e.g., the 2-parameter logistic [2PL] model), (b) combine the interdependent items to form a polytomous item and apply a polytomous IRT model (e.g., the graded response model [GRM]), and (c) apply a model that explicitly takes into account the dependence at the item level (e.g., the testlet response theory [TRT] model). In this study, a simulation was conducted to compare the performance of these 3 approaches on number-correct score equating when degrees of testlet effect were manipulated. The traditional equipercentile method was used as an evaluation baseline. The results show that the 2PL and the TRT approaches produce comparable results that more closely agree with the results of the equipercentile method than the GRM does. And the number-correct equating using the 2PL is robust to the violation of local item independence.
Citation	Cao, Y., Lu, R., & Tao, W. (2014). *Effect of item response theory (IRT) model selection on testlet-based test equating* (Research Report No. RR-14-19). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12017
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12017

Title	Demographically Adjusted Groups for Equating Test Scores
Abstract	In this study, I investigated 2 procedures intended to create test-taker groups of equal ability by poststratifying on a composite variable created from demographic information. In one procedure, the stratifying variable was the composite variable that best predicted the test score. In the other procedure, the stratifying variable was the composite that best indicated group membership (i.e., the propensity score). Applied to 2 groups taking the same test at different administrations, the composite that best predicted the test score reduced the ability difference by about two thirds; the composite that best indicated group membership reduced the ability difference by about half. Prescreening the predictor variables did not improve the performance of either procedure.
Citation	Livingston, S. A. (2014). *Demographically adjusted groups for equating test scores* (Research Report No. RR-14-30). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12030
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12030

Title	The Invariance of Latent and Observed Linking Functions in the Presence of Multiple Latent Test-Taker Dimensions
Abstract	This study examines linking relationships among latent test scores and how these latent linking relationships relate to observed-score linkings. Equations are used to describe the effects of correlation between underlying latent dimensions and the similarity or dissimilarity of test composition on linking functions among latent test scores. These equations describing relationships among latent test scores are used to model the results obtained from a previous simulation study, which illustrated that if the two tests have parallel structure then the linking relationship between their observed scores is subpopulation invariant regardless of the correlations between the underlying latent dimensions. The equations also model the effect that the degree of correlation between the latent dimensions has on equatability as the structure departs from parallelism.
Citation	Dorans, N. J., Lin, P., Wang, W., & Yao, L. (2014). *The invariance of latent and observed linking functions in the presence of multiple latent test-taker dimensions* (Research Report No. RR-14-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12041
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12041

Title	Use of Jackknifing to Evaluate Effects of Anchor Item Selection on Equating With the Nonequivalent Groups With Anchor Test (NEAT) Design
Abstract	In this study, we apply jackknifing to anchor items to evaluate the impact of anchor selection on equating stability. In an ideal world, the choice of anchor items should have little impact on equating results. When this ideal does not correspond to reality, selection of anchor items can strongly influence equating results. This influence does not disappear even if large examinee samples are present. Consequently, it provides a major hazard in practical use of equating. Although the effect of anchor selection does not disappear with increasing sample size, it is reasonable to expect smaller effects with test anchors with more items. To illustrate results, two examples of real equating data were evaluated using two classical equating methods. The results show that rather large effects may be associated with sampling of anchor items.
Citation	Lu, R., Haberman, S., Guo, H., & Liu, J. (2015). *Use of jackknifing to evaluate effects of anchor item selection on equating with the nonequivalent groups with anchor test (NEAT) design* (Research Report No. RR-15-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12056
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12056

Title	SARA Reading Components Tests, RISE Forms: Technical Adequacy and Test Design, 2nd Edition
Abstract	This technical report describes the conceptual foundation and measurement properties of the Reading Inventory and Scholastic Evaluation (RISE). The RISE is a 6-subtest, Web-administered reading skills components battery. The theoretical and empirical foundations of each subtest in the battery are reviewed, as well as item designs. The results included in this report feature a vertical extension of the RISE to span Grades 5–10, psychometric analysis of parallel forms of each subtest, results of item response theory (IRT) scaling studies for each of the subtests across the entire grade span, and evaluation of differential item functioning (DIF) for gender and race/ethnicity.
Citation	Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). *SARA reading components tests, RISE forms: Technical adequacy and test design, 2nd edition* (Research Report No. RR-15-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12076
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12076

Title	An Evaluation of the Single-Group Growth Model as an Alternative to Common-Item Equating
Abstract	As an alternative to common-item equating when common items do not function as expected, the single-group growth model (SGGM) scaling uses common examinees or repeaters to link test scores on different forms. The SGGM scaling assumes that, for repeaters taking adjacent administrations, the conditional distribution of scale scores in later administration, given the scale score in earlier administration, should generalize from previous repeaters to the repeaters taking the current administration. The current repeaters' scale score distribution is estimated from their earlier scale score distribution. The SGGM scaling first uses previous repeaters' data to estimate the conditional distribution of their later scale scores, given their earlier scale scores. Then the repeaters taking both the current and previous administrations are identified, and their scale score distribution on the current form is estimated based on their previous scale score distribution and the estimated conditional distribution. Finally, a single-group equipercentile equating is performed between the current-form repeaters' observed raw score distribution and their estimated scale score distribution to obtain the raw-to-scale score conversion. This study evaluated the SGGM scaling performance using the common-item equating results for a language test as the criterion. The study found that the raw-to-scale conversions based on SGGM scaling differed from those based on common-item equating. However, the SGGM scaling results did not show a systematic bias in either the average or the variability of examinees' scale scores.
Citation	Wei, Y., & Morgan, R. (2016). *An evaluation of the single-group growth model (SGGM) as an alternative to common-item equating* (Research Report No. RR-16-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12087
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12087

Title	Linking Composite Scores: Effects of Anchor Test Length and Content Representativeness
Abstract	The nonequivalent groups with anchor test (NEAT) design is frequently used in test score equating or linking. One important assumption of the NEAT design is that the anchor test is a miniversion of the 2 tests to be equated/linked. When the content of the 2 tests is different, it is not possible for the anchor test to be adequately representative of both tests. Lin and Dorans conducted a simulation study in 2010 to investigate the effect of content representativeness of the anchor test on linking via different linking methods when the 2 tests are nonparallel in content structure in the unique case where the groups are equivalent. The current study extends the Lin and Dorans study to the case with nonequivalent group data. Specifically, the current study investigates the impact of content representativeness and length of anchor test on linking when the 2 tests are multidimensional and nonparallel in content structure. The NEAT design was employed. The linking results from 3 classic linear equating methods—Levine observed score, Tucker equating, and chained linear—were examined. The results from the study indicated that equating the tests with different structure should be avoided. For equatings with anchor test, additional bias is likely to be introduced by using an inadequate anchor test.
Citation	Lin, P., Dorans, N., & Weeks, J. (2016). *Linking composite scores: Effects of anchor test length and content representativeness* (Research Report No. RR-16-36). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12122
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12122

Title	The Pseudo-Equivalent Groups Approach as an Alternative to Common-Item Equating
Abstract	The purpose of this study was to evaluate the effectiveness of linking test scores by using test takers' background data to form pseudo-equivalent groups (PEG) of test takers. Using 4 operational test forms that each included 100 items and were taken by more than 30,000 test takers, we created 2 half-length research forms that had either 20 (strong anchor) or 10 (weak anchor) items in common. Because the 2 research forms were assembled from a single form that had been administered in a large-scale operational testing setting, we obtained the direct equating function between the 2 research forms through the single-group design and treated it as a criterion or true equating function between the 2 research forms. We equated the 2 research forms in a common-item design using the poststratification equipercentile (PSE) and chained equipercentile (CHEQ) methods, and then compared the common-item results to the results derived from the PEG linking. Because the new and reference groups differed substantially in ability, by study design, the CHEQ method produced more accurate results than did the PSE method in both the strong and weak anchor conditions. CHEQ using 10 common items was as effective as PSE using 20 common items. PSE using 10 common items produced the least accurate results among the five methods. The PEG linking produced more accurate results compared to the PSE method using the weak anchor.
Citation	Kim, S. & Lu, R. (2018). *The pseudo-equivalent groups approach as an alternative to common-item equating* (Research Report No. RR-18-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12195
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12195

Title	A Simulation Study to Compare Nonequivalent Groups With Anchor Test Equating and Pseudo-Equivalent Group Linking
Abstract	In this paper we compare the newly developed pseudo-equivalent groups (PEG) linking method with the linking methods based on the traditional nonequivalent groups with anchor test (NEAT) design and illustrate how to use the PEG methods under imperfect equating conditions. To do this, we proposed a new method that combines the features of PEG linking and NEAT equating (referred as PEGAT) and compared it with NEAT and PEG. PEG mainly uses test takers' background variables to create PEG and then links scores on different forms whereas NEAT equating adjusts group differences in ability through the anchor test scores. The proposed method, PEGAT, uses background variables and anchor scores to adjust group ability differences. Using simulated data, these 3 linking methods were compared in 2 equating scenarios: small and large group difference in ability. The simulation design was based on real data on a test in operation. The test scores and the background variables were assumed to have a multivariate multinomial distribution. A log linear model was used to manipulate and produce simulated data. For PEG and PEGAT linking, 3 different sets of background variables were manipulated to study the impact of correlation strength of background variables to the total scores. Our results showed that NEAT linking outperformed PEG linking when the group ability difference was large, but that NEAT linking could be improved by incorporating the PEG adjustment procedure based on background variables. When the groups were similar in ability, PEG linking produced comparable results to NEAT. This finding justifies the use of PEG linking when a good anchor test is not available as well as the use of PEGAT when a good anchor is available but needs to be strengthened by background variables.
Citation	Lu, R., & Guo, H. (2018). *A simulation study to compare nonequivalent groups with anchor testing and pseudo-equivalent group linking* (Research Report No. RR-18-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12196
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12196

Title	Grouping Effects on Jackknifed Variance Estimation for Item Response Theory Scaling and Equating With Cluster-Based Assessment Data
Abstract	Educational assessment data are often collected from a set of test centers across various geographic regions, and therefore the data samples contain clusters. Such cluster-based data may result in clustering effects in variance estimation. However, in many grouped jackknife variance estimation applications, jackknife groups are often formed by a random grouping method that ignores the cluster structures of the data. In this study, we constructed both random and cluster-based jackknife groups for data known to have cluster structures and compared the jackknifed standard errors, yielded by two different grouping methods, of item response theory (IRT) scaling coefficient estimates and equated scores. Three independent data samples from an international test of English were used for the study. The cluster-based jackknife group results showed relatively larger jackknifed standard errors of scaling coefficient estimates and scale scores than the results of the random jackknife groups for all three data samples. For cluster-based assessment data, the cluster-based jackknife approach provides a more appropriate way to estimate the standard errors of the parameters of IRT calibration, scaling, and equating analyses.
Citation	Wang, L., Qian, J., & Lee, Y.-H. (2018). *Grouping effects on jackknifed variance estimation for item response theory scaling and equating with cluster-based assessment data* (Research Report No. RR-18-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12204
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12204

Title	Different Methods of Adjusting for Form Difficulty Under the Rasch Model: Impact on Consistency of Assessment Results
Abstract	When using the Rasch model, equating with a nonequivalent groups anchor test design is commonly achieved by adjustment of new form item difficulty using an additive equating constant. Using simulated 5-year data, this report compares 4 approaches to calculating the equating constants and the subsequent impact on equating results. The 4 approaches are mean difference, mean difference with outlier removal using the 0.3 logit rule, mean difference with robust z statistic, and the information-weighted mean difference. Factors studied included sample size, anchor test length, percentage of anchor items displaying outlier behavior, and the distribution of test item difficulty relative to examine ability. The results indicated that the mean difference and information-weighted mean difference methods performed similarly across all conditions. In addition, with larger sample sizes, the mean difference with 0.3 logit method performed similarly to these 2 methods. The mean difference with robust z method performed most differently from the other three methods of calculating the equating constant. This method removed a large percentage of the anchor items compared to the mean difference with 0.3 logit method but seemed to produce the most stable trend in performance classification across the 5 years, particularly when sample sizes were large.
Citation	Manna, V. F., & Gu, L. (2019). *Different methods of adjusting for form difficulty under the Rasch model: Impact on consistency of assessment results* (Research Report No. RR-19-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12244
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12244

Title	Error Variance in Common Population Linking Bridge Studies
Abstract	When an assessment undergoes changes to the administration or instrument, bridge studies are typically used to try to ensure comparability of scores before and after the change. Among the most common and powerful is the common population linking design, with the use of a linear transformation to link scores to the metric of the original assessment. In the common population linking design, randomly equivalent samples receive the new and previous administration or instrument. However, conventional procedures to estimate error variances are not appropriate for scores linked in a bridge study, because the procedures neglect variance due to linking. A convenient approach is to estimate a variance component associated with the linking to add to the conventionally estimated error variance. Equations for the variance components in this approach are derived, and the approximations inherently made in this approach are shown and discussed. Exact error variances of linked scores, accounting for both conventional sources of variance (e.g., sampling) and linking variance together, are derived and discussed. The consequences of how linking changes how certain errors are related is considered mathematically. Specifically, the impacts of linking on the error variance for the comparison of two linked estimates (e.g., comparing the mean score of boys to the mean score of girls, after linking), for the comparison of scores across the two samples (e.g., comparing the mean score of boys in the new administration or instrument to the mean score of boys in the old administration or instrument), and for aggregating scores across the two samples (e.g., the mean score of boys across both administrations or instruments) are derived and discussed. Finally, general methods to account for error variance in bridge studies by simultaneously accounting for both conventional and linking sources of error are recommended.
Citation	Jewsbury, P. A. (2019). *Error variance in common population linking bridge studies* (Research Report No. RR-19-42). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12279
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12279

Title	Uncommon Measures Revisited
Abstract	This report, which is based on an invited presentation given at the 2015 meeting of the Association of Test Publishers, is a response to the continuing proliferation of scale linking studies that have occurred since the publication of Uncommon Measures in 1999. The report has four parts. First, I restate the conclusions made in Uncommon Measures about linking the scales of state assessments to the National Assessment of Educational Progress scale and summarize points made by Thissen with respect to such linkages. Then I reiterate the important role played by the features of testing situations on the type of linkages made by Kolen and note how these features interact with the taxonomy of score linkages provided by Holland and Dorans. Next, I summarize findings from a 2010 National Council on Measurement in Education symposium that described the linkage studies conducted in 2008 to update the concordances between the 2005 version of the SAT® test and the ACT, and I discuss their implications for linking score scales in general. Finally, I offer some concluding advice pertaining to linkages in general.
Citation	Dorans, N. J. (2020). *Uncommon measures revisited* (Research Report No. RR-20-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12287
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12287

Title	Effect of Statistically Matching Equating Samples for Common-Item Equating
Abstract	This study evaluated the impact of subgroup weighting for equating through a common-item anchor. We used data from a single test form to create two research forms for which the equating relationship was known. The results showed that equating was most accurate when the new form and reference form samples were weighted to be similar to the target population. When the target population was a combination of the two equating samples and one sample was weighted to be similar to the other, the equating was less accurate but still much more accurate than equating with unweighted samples.
Citation	Lu, R., & Kim, S. (2021*). Effect of statistically matching equating samples for common-item equating* (Research Report No. RR-21-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12313
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12313

Title	Assessing Mode Effects of At-Home Testing Without a Randomized Trial
Abstract	In this investigation, we used real data to assess potential differential effects associated with taking a test in a test center (TC) versus testing at home using remote proctoring (RP). We used a pseudo-equivalent groups (PEG) approach to examine group equivalence at the item level and the total score level. If our assumption holds that the PEG approach removes between-group ability differences (as measured by the test) reasonably well, then a plausible explanation for any systematic differences in performance between TC and RP groups that remain after applying the PEG approach would be the operation of test mode effects. At the item level, we compared item difficulties estimated using the PEG approach (i.e., adjusting only for ability differences between groups) to those estimated via delta equating (i.e., adjusting for any systematic differences between groups). All tests used in this investigation showed small, nonsystematic differences, providing evidence of trivial effects associated with at-home testing. At the total score level, we linked the RP group scores to the TC group scores after adjusting for group differences using demographic covariates. We then compared the resulting RP group conversion to the original TC group conversion (the criterion in this study). The magnitude of differences between the RP conversion and the TC conversion was small, leading to the same pass/fail decision for most RP examinees. The present analyses seem to suggest little to no mode effects for the tests used in this investigation.
Citation	Kim, S., & Walker, M. (2021). *Assessing mode effects of at-home testing without a randomized trial* (Research Report No. RR-21-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12323
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12323

Title	Comparisons Among Approaches to Link Tests Using Random Samples Selected Under Suboptimal Conditions
Abstract	Equating the scores from different forms of a test requires collecting data that link the forms. Problems arise when the test forms to be linked are given to groups that are not equivalent and the forms share no common items by which to measure or adjust for this group nonequivalence. We compared three approaches to adjusting for group nonequivalence in a situation where not only is randomization questionable, but the number of common items is small. Group adjustment through either subgroup weighting, a weak anchor, or a mix of both was evaluated in terms of linking accuracy using a resampling approach. We used data from a single test form to create two research forms for which the equating relationship was known. The results showed that both subgroup weighting and weak anchor approaches produced nearly equivalent linking results when group equivalence was not met. Direct (random groups) linking methods produced the least accurate result due to nontrivial bias. Use of subgroup weighting and linking using the anchor test only marginally improved linking accuracy compared to using the weak anchor alone when the degree of group nonequivalence was small.
Citation	Kim, S., & Walker, M. (2021). *Comparisons among approaches to link tests using random samples selected under suboptimal conditions* (Research Report No. RR-21-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12328
Link	https://onlinelibrary.wiley.com/doi/10.1002/ets2.12328

Selected Research Reports: ACT

Title	Evaluating the Effects of Differences in Group Abilities on the Tucker and the Levine Observed-Score Methods for Common-Item Nonequivalent Groups Equating
Abstract	The most critical feature of a common-item nonequivalent groups equating design is that the average score difference between the new and old groups can be accurately decomposed into a group ability difference and a form difficulty difference. Two widely used observed-score linear equating methods, the Tucker and the Levine observed-score methods, have different statistical assumptions when decomposing the score difference. Variation in the decomposition of group ability and form difficulty differences can affect the equating results. This study confirmed previous findings in the literature that when form and group differences are small, both equating methods produce similar results. When the group ability difference is large, however, the Levine observed-score method produces more accurate equating results than the Tucker method. The results indicated that the Levine observed-score method not only decomposes form and group differences more accurately, but also yields smaller unweighted absolute equating differences and average weighted root mean square differences. This study showed that the Levine observed-score method is also robust to the form difference.
Citation	Chen, H., Cui, Z., Zhu, R., & Gao, X. (2010). *Evaluating the effects of differences in group abilities on the Tucker and the Levine observed-score methods for common-item nonequivalent groups equating* (Research Report 2010-1). Iowa City, IA: American College Testing Program.
Link	https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2010-1.pdf

Title	A Comparison of Four Linear Equating Methods for the Common-Item Nonequivalent Groups Design Using Simulation Methods
Abstract	This paper investigates four methods of linear equating under the common item nonequivalent groups design. Three of the methods are well known: Tucker, Angoff-Levine, and Congeneric-Levine. A fourth method is presented as a variant of the Congeneric-Levine method. Using simulation data generated from the three-parameter logistic IRT model we compare the accuracy of the four methods under a variety of conditions involving group differences between the old and new groups. The sampling properties of the methods' parameter estimates are also investigated. The results indicate that the Tucker method is less accurate than the other three methods when group differences exist, especially when sample size is large (800). However, the Tucker method's gamma has the smallest sampling error, especially when sample size is small.
Citation	Topczewski, A., Cui, Z., Woodruff, D., Chen, H., & Fang, Y. (2013). *A comparison of four linear equating methods for the common-item nonequivalent groups design using simulation methods* (Research Report 2013-2). Iowa City, IA: American College Testing Program.
Link	https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2013-2.pdf

Title	A Comparison of Three Methods for Computing Scale Score Conditional Standard Errors of Measurement
Abstract	Professional standards for educational testing recommend that both the overall standard error of measurement and the conditional standard error of measurement (CSEM) be computed on the score scale used to report scores to examinees. Several methods have been developed to compute scale score CSEMs. This paper compares three methods, based on classical test theory, item response theory, and the four-parameter beta compound binomial model. The three methods are compared using data from a single form of the ACT® College Readiness Assessment. The results indicate that all three methods produce comparable results.
Citation	Woodruff, D., Traynor, A., Cui, Z., & Fang, Y. (2013). *A comparison of three methods for computing scale score conditional standard errors of measurement* (Research Report 2013-7). Iowa City, IA: American College Testing Program.
Link	https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2013-7.pdf

Title	Concordance of ACT Aspire and PreACT/ACT Test Scores
Abstract	(N/A)
Citation	Allen, J., & Tao, W. (2020). *Concordance of ACT Aspire and PreACT/ACT test scores* (ACT Research & Policy Technical Brief). Iowa City, IA: American College Testing Program.
Link	https://www.act.org/content/dam/act/unsecured/documents/R1816-aspire-act-preact-concordance.pdf

Title	An Investigation of ACT Equating Stability
Abstract	Equating is the statistical procedure that adjusts scores from a given ACT® test form to make them interchangeable with scores from other ACT test forms, which may differ slightly in difficulty. Stability is an important property of equating results because it ensures that score meaning is consistent over time and independent of the examinee samples used for equating. ACT regularly conducts "stability check" analyses by reequating forms that were equated previously. This report describes one such analysis conducted after the February 2020 ACT administration. Overall, results indicated that differences in equating results were within the range of what would be expected due to random sampling error. Therefore, this study provides evidence supporting the stability of ACT equating results.
Citation	Li, D. (2020). *An investigation of ACT equating stability* (ACT Research & Policy Technical Brief). Iowa City, IA: American College Testing Program.
Link	https://www.act.org/content/dam/act/unsecured/documents/R1837-ACT-equating-stability-2020-11.pdf

Title	Finding Stability: Comparing Methods for Detecting Unstable Item Parameters in IRT Equating
Abstract	In IRT-based common item equating, instability in common item parameters can introduce error into IRT scale transformations, subsequent equating results, and, ultimately, examinee scores. This study compared five methods of identifying items with significant parameter drift. Rather than detecting simulated parameter drift like many prior studies, this study used expected equating results as evaluation criteria, which was possible due to the operational use random groups equipercentile equating with an anchor form. Results indicated that two methods produced similarly low equating error while eliminating relatively few items from the common item set. The first was ACT's current practice of flagging items with outlier parameter estimates based on historical distributions. The second was the Delta method, which flags items when transformed proportion correct values are significantly different from expectations.
Citation	Steedle, J. T. (2022). *Finding stability: Comparing methods for detecting unstable item parameters in IRT equating* (ACT Research Technical Brief). Iowa City, IA: American College Testing Program.
Link	https://www.act.org/content/dam/act/unsecured/documents/2022/R2153-Finding-Stability-Detecting-Unstable-Item-Parameters-in-IRT-Equating-2022-03.pdf

Selected Research Reports: College Board

Title	Exploring Equity Properties in Equating Using AP® Examinations
Abstract	In almost all high-stakes testing programs, test equating is necessary to ensure that test scores across multiple test administrations are equivalent and can be used interchangeably. Test equating becomes even more challenging in mixed-format tests, such as Advanced Placement Program® (AP®) Exams, that contain both multiple-choice and constructed response items. This report examines (1) the performance of various equating methods in terms of first- and second-order equity properties using mixed-format tests; (2) the effect of underlying psychometric models on the assessment of the performance of the equating methods; and (3) the relationship between reliability and equity properties in equating. Three AP Exams (Biology, English Language and Composition, and French Language and Culture) were analyzed with the common-item, nonequivalent-groups design. The 11 equating methods were analyzed, and the results were obtained and compared based upon two different psychometric model frameworks: the two-parameter beta binomial and item-response theory (IRT). In general, the results showed that the performance of various equating methods in terms of equity properties depended on the psychometric model assumed. Furthermore, this report provides empirical evidence that the magnitude of reliability plays a role in achieving the equity properties for the various equating methods.
Citation	Lee, E., Lee, W.-C., & Brennan, R. L. (2012). *Exploring equity properties in equating using AP® examinations* (Research Report 2012-4). New York: College Board.
Link	https://files.eric.ed.gov/fulltext/ED561043.pdf

Title	Scaling for the SAT Suite of Assessments
Abstract	(N/A)
Citation	College Board. (2017). *Scaling for the SAT Suite of Assessments*. New York: College Board.
Link	https://satsuite.collegeboard.org/media/pdf/scaling-sat-suite-assessments.pdf

Selected Research Reports: Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa

Monographs	Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating
Citation and Link	Kolen, M. J., & Lee, W.-C. (Eds.). (2011). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1)*. (CASMA Monograph Number 2.1). Iowa City, IA: CASMA, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.1.pdf
	Kolen, M. J., & Lee, W.-C. (Eds.). (2012). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2)*. (CASMA Monograph Number 2.2). Iowa City, IA: CASMA, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.2.pdf
	Kolen, M. J., & Lee, W.-C. (Eds.). (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 3)*. (CASMA Monograph Number 2.3). Iowa City, IA: CASMA, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.3.pdf
	Kolen, M. J., & Lee, W.-C. (Eds.). (2016). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 4)*. (CASMA Monograph Number 2.4). Iowa City, IA: CASMA, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.4.pdf
	Kolen, M. J., & Lee, W.-C. (Eds.). (2018). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 5)*. (CASMA Monograph Number 2.5). Iowa City, IA: CASMA, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.5.pdf

Title	First-order and Second-order Equity in Equating
Abstract	(N/A)
Citation	Brennan, R. L. (2010). *First-order and second-order equity in equating*. (CASMA Research Report No. 30). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-30.pdf

Title	Assessing Equating Results Based on First-order and Second-order Equity
Abstract	(N/A)
Citation	Lee, E. J., Lee, W.-C., & Brennan, R. L. (2010). *Assessing equating results based on first-order and second-order equity*. (CASMA Research Report No. 31). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-31.pdf

Title	Effects of the Number of Common Items on Equating Precision and Estimates of the Lower Bound to the Number of Common Items Needed
Abstract	(N/A)
Citation	Zhang, M., & Kolen, M. J. (2013). *Effects of the number of common items on equating precision and estimates of the lower bound to the number of common items needed*. (CASMA Research Report No. 37). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-37.pdf

Title	Equating Multidimensional Tests under a Random Groups Design: A Comparison of Various Equating Procedures
Abstract	(N/A)
Citation	Lee, E., Lee, W.-C., & Brennan, R. L. (2014). *Equating multidimensional tests under a random groups design: a comparison of various equating procedures*. (CASMA Research Report No. 40). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-40.pdf

Title	Notes about Partial Derivatives for Analytic Standard Errors of Levin-observed Equating with an External Anchor
Abstract	(N/A)
Citation	Kim, H. J., & Brennan, R. L. (2015). *Notes about partial derivatives for analytic standard errors of Levin-observed equating with an external anchor*. (CASMA Technical Notes No. 8). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-technical-report-8.pdf

Title	A Note on Sample Size for Alternative Random Groups Equating Designs
Abstract	(N/A)
Citation	Kolen, M. J. (2015). *A note on sample size for alternative random groups equating designs*. (CASMA Technical Notes No. 7). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-technical-report-7.pdf

Title	Equating with Bivariate Log-linear Presmoothing under the Common-item Nonequivalent Groups Design: Structural Zeros and Their Implications
Abstract	(N/A)
Citation	Kim, H. J., Brennan, R. L., & Lee, W.-C. (2015). *Equating with bivariate log-linear presmoothing under the common-item nonequivalent groups design: Structural zeros and their implications*. (CASMA Research Report No. 43). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-43.pdf

Title	Subscore Equating and Reporting
Abstract	(N/A)
Citation	Lim, E., & Lee, W.-C. (2016). *Subscore equating and reporting*. (CASMA Research Report No. 47). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-47.pdf

Title	Similarities Between Equated Equivalents Using Presmoothing and Postsmoothing
Abstract	(N/A)
Citation	Kim, H. J., Brennan, R. L., & Lee, W.-C. (2016). *Similarities between equated equivalents using presmoothing and postsmoothing*. (CASMA Research Report No. 48). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-48.pdf

Title	Multiple Group IRT Fixed-Parameter Estimation for Maintaining an Established Ability Scale
Abstract	(N/A)
Citation	Kim, S., & Kolen, M. (2016). *Multiple group IRT fixed-parameter estimation for maintaining an established ability scale*. (CASMA Research Report No. 49). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-49.pdf

Title	A Statistical Criterion to Assess Fitness of Cubic-Spline Postsmoothing
Abstract	(N/A)
Citation	Kim, H. J., Brennan, R. L., & Lee, W.-C. (2017). *A statistical criterion to assess fitness of cubic-spline postsmoothing*. (CASMA Research Report No. 52). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-52.pdf

Title	Impact of Degrees of Postsmoothing on Long-Term Equated Scale Score Accuracy
Abstract	(N/A)
Citation	Kim, S., Kim, Y., & Moses, T. (2020). *Impact of degrees of postsmoothing on long-term equated scale score accuracy*. (CASMA Research Report No. 54). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Link	https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-54.pdf