© IIJP 2025 | Volume 1, Issue 4, pp. 312-327 | ISSN: 2636-4484

ADVANCING FAIRNESS IN 21ST CENTURY TESTING: A SYNTHESIS OF TEST EQUATING STRATEGIES AND OUTCOMES

Ogunsakin, Isaac Bamikole
Department of Educational Foundations and Counselling, Obafemi Awolowo University, Ile-Ife, Osun State Nigeria
sakinbamikole2019@gmail.com/08039215009

ARTICLE INFO
Article No.: 091
Accepted Date: 7/11/2025
Published Date: 17/11/2025
Type: Research

ABSTRACT
Ensuring fairness and validity in high-stakes educational and psychological assessments requires accurate comparison of scores across multiple test forms. This paper provides a comprehensive examination of test equating, highlighting its theoretical foundations, methodologies, and practical applications. Drawing on Classical Test Theory (CTT) and Item Response Theory (IRT), the study reviews linear, equipercentile, mean, chain, Haebara, and Stocking-Lord equating methods, as well as horizontal and vertical equating designs. Emphasis is placed on the critical role of equating in maintaining score comparability, addressing item parameter drift, supporting high-stakes decisions, and enhancing test security, including mitigating item exposure and cheating. Recent advances in propensity score-based equating and automated item generation are also discussed as innovative solutions for contemporary testing challenges. The synthesis identifies gaps in empirical research regarding the application of equating methods in digital and open-testing environments. The paper underscores equating as not only a statistical procedure but also a tool for fairness, ensuring that examinee performance is measured accurately and consistently across test forms and administrations.

Keywords: Test, Test Equating, Classical Test Theory and Item Response Theory

Introduction

Advancing fairness in testing is a critical focus in the 21st century, particularly in the context of high-stakes assessments that significantly influence individual and societal outcomes. Test equating plays a pivotal role in ensuring the reliability, validity, and comparability of scores across different test forms or administrations (Kolen & Brennan, 2014). This process addresses potential disparities introduced by varying test conditions, enabling examinees to be evaluated on a common scale regardless of the specific test form administered (Sireci, 2005). As testing programs increasingly adopt open testing models and digital platforms, the need for robust equating methodologies has intensified to uphold fairness in a rapidly evolving educational and professional landscape (von Davier et al., 2024). Recent studies further emphasize the importance of equating in maintaining test fairness and validity across diverse testing contexts (Wallin & Wiberg, 2024; Tavakol et al., 2024).

Effective equating ensures that test scores accurately reflect the knowledge or ability of examinees rather than artifacts of the specific test form or administration conditions. From my perspective, it is not enough for institutions to simply adopt standard equating procedures; there must be a conscious effort to evaluate how these methods function in practice and whether they truly promote fairness. I argue that advances in computational psychometrics and data-driven equating techniques provide unprecedented opportunities to refine score comparability, especially in large-scale digital assessments. In my view, equating is more than a technical procedure it is a tool for social justice, helping to reduce biases that may disproportionately affect certain groups and ensuring that all examinees are evaluated on a level playing field. Therefore, I believe that a rigorous, reflective approach to implementing and monitoring equating methodologies is essential to sustain credibility, fairness, and public confidence in contemporary assessment systems.

In educational and psychological measurement, a test consisting of a set of items is typically administered to a sample of subjects to make inference on the latent variables underlying the response process (Valentina, Marie and Mariagiulia 2017) the authors further explained that latent variables are not directly observed but are rather inferred through a statistical model from the observed, directly measured, item responses. Statistical models that aim to explain observed variables in terms of latent variables are called latent variable models. Latent variable models are used in many disciplines, including psychology, economics and the social sciences. Examples of latent variables in the field of economics include quality of life and happiness; in an educational context a typical latent variable is the examinee's ability on a specific subject (e.g., mathematics). A typical situation in these fields is that different tests are used to measure the same latent variable.

Test score equating is essential for comparing scores obtained from different forms of a test, ensuring that results are interpreted on a consistent scale regardless of the specific version taken (Kolen & Brennan, 2014; González & Wiberg, 2017). There are three key reasons why multiple forms of a test are often required, all of which make equating a necessary process.

The first reason is the need for security, particularly in high-stakes testing environments. In many educational and professional contexts, test outcomes have significant consequences, such as granting a license to practice a profession, admitting a candidate into a university program, or awarding academic credit. Repeated use of the same test form can lead to item exposure, which compromises fairness and undermines the integrity of the assessment. To protect against this, different forms of a test must be developed, and equating is used to ensure scores remain comparable (Ryan & Brockmann, 2009; Issayeva, 2021).

The second reason relates to the growing movement towards open testing. Many testing programs now release test items to the public to promote transparency and accountability (Brauneis & Goodman, 2018). However, once items are publicly available, they can no longer be reused without giving certain test-takers an unfair advantage. As a result, new test forms must be created regularly, making equating essential to ensure consistency in scoring across different versions (Ryan & Brockmann, 2009).

The third reason stems from the natural evolution of test content over time. As educational standards shift and the knowledge base within disciplines grows or changes, test items must be updated to reflect current expectations. This ongoing need to revise and replace outdated items leads to the development of new test forms. Equating is necessary in this context to maintain the validity and comparability of scores across different test administrations, ensuring that all examinees are assessed fairly, regardless of when or which form of the test they take (Issayeva, 2021).

What is Test Equating?

Test equating is a statistical process used to adjust scores on different test forms so that they can be used interchangeably. This process is essential for ensuring fairness, comparability, and validity across multiple administrations of a test. In large-scale assessments, multiple test forms are often necessary to prevent item overexposure, which could compromise the security and integrity of the examination. Without equating, scores from different forms could not be compared meaningfully, as slight variations in difficulty or item content could result in inaccurate interpretations of examinee ability. Recent research has emphasized that equating is particularly important when dealing with diverse populations, multiple administrations, and international assessments, where contextual and demographic differences can influence performance. For instance, Uzun (2025) examined the PISA 2018 mathematics subtest and demonstrated how differential item functioning (DIF) across countries can affect equating results, highlighting the necessity of robust statistical procedures to maintain score comparability.

` The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) define equating as the process of placing scores from two or more parallel tests on a common scale. This definition underscores the purpose of equating: to ensure that scores obtained from different test forms are comparable, meaningful, and interpretable in the same way. In practice, equating involves statistical adjustments that account for differences in test

difficulty, item characteristics, and examinee populations. Leôncio et al. (2022) highlight several equating methodologies, including classical observed-score equating and item response theory (IRT)–based approaches, which allow testing programs to produce comparable score scales even when test forms differ slightly in difficulty or format. This ensures that decisions based on test scores such as admissions, certifications, or licensure is fair and evidence-based.

Equating is particularly crucial in high-stakes testing environments, where the consequences of score misinterpretation can be significant. For example, professional licensure exams, university entrance tests, and other high-stakes assessments require precise score comparability to maintain the credibility and reliability of the evaluation system. To illustrate, imagine two thermometers—one digital and one mercury-based used to measure the same individual's body temperature. While each thermometer may display slightly different readings due to calibration differences, both measure the same underlying condition: the person's actual temperature. By applying a calibration adjustment, the readings can be aligned to accurately reflect the same value. Similarly, equating aligns scores from different test forms so that they reflect the same level of ability, despite minor differences in difficulty or format. This process allows educators, policymakers, and test developers to make informed and equitable decisions based on comparable data, enhancing the overall fairness and effectiveness of educational and professional assessment systems.

Equating is often considered the most stringent process for creating comparable scores. Recent studies have emphasized its role in ensuring score comparability across different test forms. For instance, a 2023 study highlighted the importance of equating in maintaining fairness and reliability in assessments (Sun, 2023). Other methods, such as calibration, concordance, statistical moderation, and prediction, are also used to convert scores from one test to another. These methods, collectively referred to as linking, do not necessarily meet the stringent assumptions required for equating. A 2025 study discussed the use of concordance statistics in evaluating treatment benefit predictors, illustrating one aspect of linking methods (Curcin, 2025). It is important to note that while all procedures for linking scores aim to produce comparable scores, only equating provides interchangeable scores. The term 'equating' is, therefore, strictly reserved for score conversions for alternate forms of a test tests that measure the same content and are built to the same specifications leading to interchangeable scores.

The chart in Figure 1.0 offers a clear and comprehensive overview of the structure and process of test equating, highlighting the different methods and designs used to compare scores across multiple test forms or administrations. From my perspective, it is fascinating how these approaches balance the technical rigor of statistical methods with practical concerns about fairness and validity. The chart distinguishes between two fundamental types of equating: Vertical Equating and Horizontal Equating, each serving distinct purposes in educational and psychological measurement. Vertical equating is applied when tests are designed for different grade levels or ability groups, allowing meaningful comparisons across tests of differing

difficulty. In contrast, horizontal equating is used when multiple forms of a test measure the same construct at the same level, ensuring that all examinees are evaluated fairly and equitably.

These approaches are anchored in two major theoretical frameworks: Classical Test Theory (CTT) and Item Response Theory (IRT). Under CTT, methods such as mean, linear, chain, and equipercentile equating adjust scores based on observed test statistics. IRT-based methods, on the other hand, work at the item level, using logistic models (1PL, 2PL, and 3PL) to provide greater precision and flexibility in score comparisons. In my view, what makes IRT particularly compelling is how it allows test developers to capture nuances in item difficulty and examinee ability, creating a more refined and equitable evaluation process. Overall, Figure 1.0 not only clarifies the technical aspects of equating but also underscores the critical role these methods play in maintaining fairness, validity, and confidence in assessment outcomes.

Furthermore, the figure emphasizes the essential role of equating designs in test equating studies. Design frameworks such as single group, counterbalanced group, and equivalent group provide structured approaches for collecting data necessary for equating, while calibration and linking methods like Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord enhance the accuracy and validity of score alignment. Generally, Figure 1.0 demonstrates that effective test equating requires both methodological rigor and theoretical grounding, integrating statistical procedures, psychometric principles, and design considerations to ensure that test scores from different forms are comparable, fair, and interpretable, thereby supporting valid decisions about examinee performance.

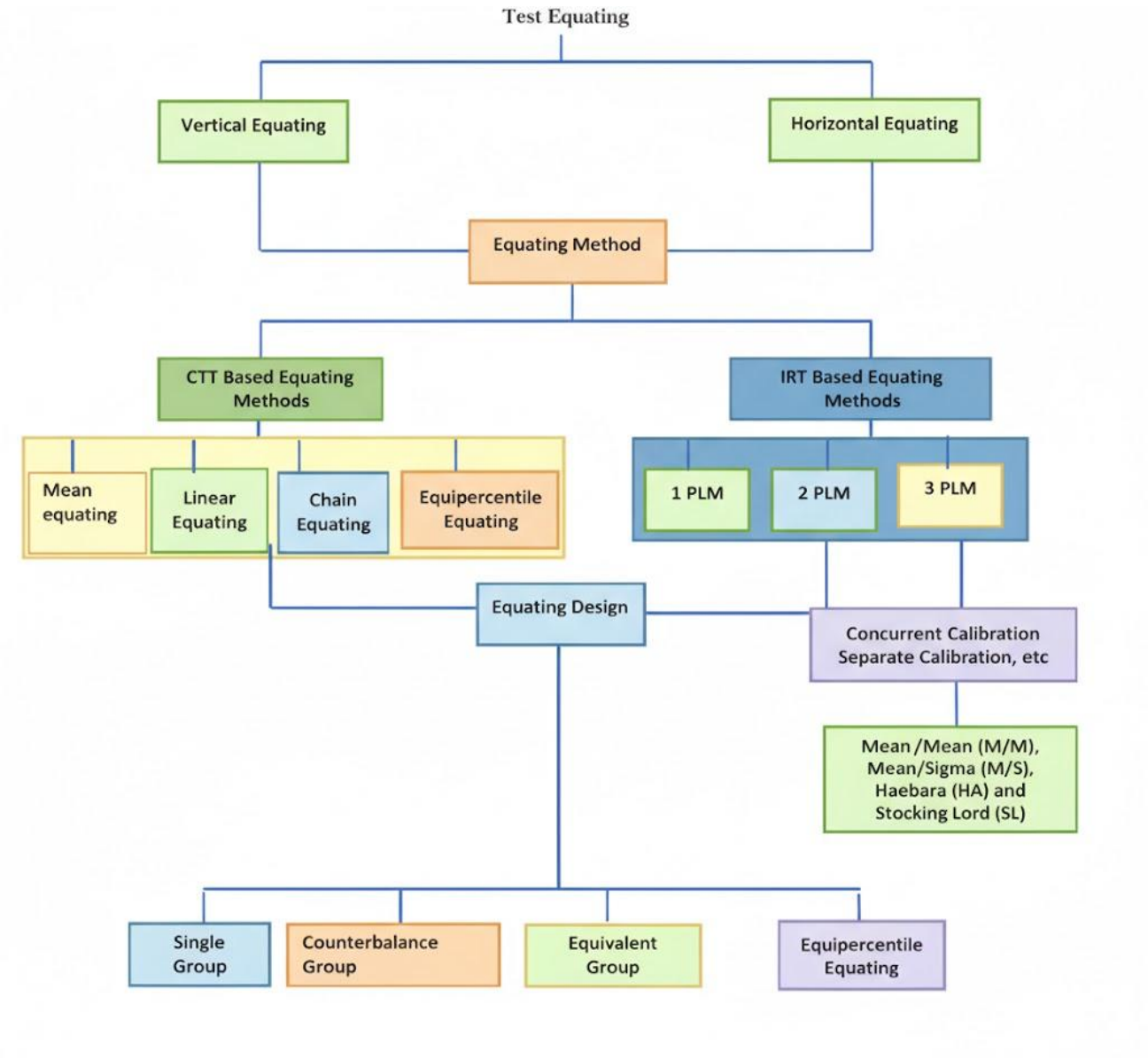

Fig: 1.0: Chart of Test Equating.
Source: Agah (2013)

Vertical and Horizontal Equating

Test score equating is a statistical process used to adjust scores on different test forms so that they can be used interchangeably. This process is essential for ensuring fairness, comparability, and validity across multiple administrations of a test. In large-scale assessments, multiple test forms are often necessary to prevent item overexposure, which could compromise the security and integrity of the examination. Without equating, scores from different forms could not be compared meaningfully, as slight variations in difficulty or item content could result in inaccurate interpretations of examinee ability.

Recent research has emphasized that equating is particularly important when dealing with diverse populations, multiple administrations, and international assessments, where contextual and demographic differences can influence performance. For instance, Uzun (2025) examined the PISA 2018 mathematics subtest and demonstrated how differential item functioning (DIF) across countries can affect equating results, highlighting the necessity of robust statistical procedures to maintain score comparability.

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) define equating as the process of placing scores from two or more parallel tests on a common scale. This definition underscores the purpose of equating: to ensure that scores obtained from different test forms are comparable, meaningful, and interpretable in the same way. In practice, equating involves statistical adjustments that account for differences in test difficulty, item characteristics, and examinee populations. Leôncio et al. (2022) highlight several equating methodologies, including classical observed-score equating and item response theory (IRT)–based approaches, which allow testing programs to produce comparable score scales even when test forms differ slightly in difficulty or format. This ensures that decisions based on test scores such as admissions, certifications, or licensure are fair and evidence-based.

Equating is particularly crucial in high-stakes testing environments, where the consequences of score misinterpretation can be significant. For example, professional licensure exams, university entrance tests, and other high-stakes assessments require precise score comparability to maintain the credibility and reliability of the evaluation system. To illustrate, imagine two thermometers—one digital and one mercury-based used to measure the same individual's body temperature. While each thermometer may display slightly different readings due to calibration differences, both measure the same underlying condition: the person's actual temperature. By applying a calibration adjustment, the readings can be aligned to accurately reflect the same value. Similarly, equating aligns scores from different test forms so that they reflect the same level of ability, despite minor differences in difficulty or format. This process allows educators, policymakers, and test developers to make informed and equitable decisions based on comparable data, enhancing the overall fairness and effectiveness of educational and professional assessment systems.

Additionally, Horizontal equating is executed between two different forms of a test (form A and form B). An example would be equating various forms of the JAMB, WAEC, NECO and NABTEB examination as they are administered across a 2 or 3 year period. It is important that the scores should be reasonably comparable across forms and time. Ths is one of the principal application of test equating examination bodies across the globe.

Clasical Test Theory Based Equating Methods

Classical Test Theory (CTT) provides foundational methods for test score equating, including linear, equipercentile, chain, and mean equating. Linear equating assumes a linear

relationship between scores and adjusts for mean and standard deviation differences (Livingston & Kim, 2009). Equipercentile equating aligns scores based on percentile ranks, avoiding linearity assumptions (Babcock, Albano, & Raymond, 2012). Chain equating uses intermediate forms when direct equating isn't possible (von Davier & Kong, 2003), while mean equating simply aligns test forms by their average scores under the assumption of parallel forms (Ryan, 2009). Each method has unique strengths and limitations depending on the test context and underlying assumptions (Livingston & Kim, 2009).

Linear Equating

Linear equating is widely utilized in both operational and experimental settings. For instance, a study by Kolen and Brennan (2014) demonstrated that linear equating effectively maintained score comparability in a high-stakes testing context when forms of a college admissions exam were equated. Their findings showed that the test forms' score distributions were sufficiently similar, which made the linear model both accurate and reliable. However, they also observed that when the score distributions were significantly different, as in some longitudinal studies, the equating results were less reliable. Empirical

evidence from this study indicated that linear equating produced acceptable levels of accuracy in large-scale assessments, where the test items across forms measured the same construct (Kolen & Brennan, 2014).

Another empirical study by Livingston (2004) highlighted that, in the context of a high-stakes licensing exam, linear equating was effective when test forms were parallel, ensuring that examinees received comparable scores. However, this study also found that small differences in difficulty levels between forms led to discrepancies in equated scores, suggesting that linear equating may not always capture subtler variations in item difficulty.

Equipercentile Equating

Equipercentile equating has been shown to be particularly useful in situations where the score distributions of test forms differ significantly (Sun & Wang, 2023; Gübeşi & Uyar, 2020). In a large-scale educational testing program, Livingston (2004) compared the performance of equipercentile equating against linear equating. The results indicated that equipercentile equating provided a better alignment of score percentiles, leading to more accurate comparability between forms with skewed or bimodal distributions (Livingston, 2004; Sun & Wang, 2023). The empirical results also showed that equipercentile equating was able to produce equated scores that were closer to actual performance levels for different student populations (Gübeşi & Uyar, 2020).

An additional study by von Davier (2017) applied equipercentile equating to equate scores on a teacher certification exam with large differences in score distributions across years. The results showed that equipercentile equating outperformed linear methods in ensuring score comparability, particularly for those at the lower end of the score distributions. These results were confirmed by subsequent validation studies, which demonstrated that equipercentile equating significantly improved fairness when test forms had different difficulty levels

Chain Equating

Chain equating, while useful in large-scale testing programs, has been found to be prone to accumulation of error, particularly when the intermediate forms are not sufficiently equated themselves. A study by von Davier (2017) assessed the efficacy of chain equating across three test forms administered to different groups of examinees. The study revealed that errors in

intermediate chains compounded as the number of test forms increased, especially when there were discrepancies in the difficulty of the forms. The empirical findings highlighted the risk of inflation or deflation of scores as the equating error propagated through the chain.

Chain equating remains a widely used method in large-scale testing programs, particularly when direct equating between two test forms is not feasible. Holland and Dorans (2006) provided a comprehensive framework for understanding and implementing chain equating within the context of non-equivalent groups with anchor test (NEAT) designs. Their work emphasized the importance of ensuring that intermediate forms used in the chain are psychometrically equivalent to the target forms to maintain the validity of the equating process. Recent studies have continued to explore the efficacy and limitations of chain equating. For instance, Aşiret (2023) examined the impact of missing data on equating methods and found that chain equating methods, when appropriately applied, can yield reliable results even in the presence of incomplete data. However, the study also highlighted that the reliability of equated scores diminishes if the linking process is flawed or if the intermediate forms are not well-constructed. These findings underscore the critical role of careful design and implementation in the chain equating process to ensure the comparability and fairness of test scores across different forms and administrations.

Mean Equating

Mean equating is often employed when test forms are relatively similar, and the primary concern is to adjust for differences in overall difficulty. Recent studies have highlighted that even minor differences in test difficulty can impact the equating process. For instance, Wyse (2018) discussed the challenges of equating Angoff standard-setting ratings with the Rasch model, emphasizing that small variations in item difficulty can lead to significant differences in equated scores. These findings underscore the importance of ensuring that test forms are truly parallel when applying mean equating methods. Empirical results from this study indicated that the mean equating method could provide a simple yet effective solution for minor variations in test difficulty, particularly when the score distributions were approximately normal.

However, subsequent studies have shown that mean equating may fail to adequately adjust for variations in score distributions across forms. For example, a study by Kolen and Brennan (2014) applied mean equating in a context where two forms of a professional certification exam showed substantial differences in score distributions. The results indicated that while mean equating adjusted for differences in overall mean scores, it did not account for differences in the variability of test scores, which led to underestimation of scores for some examinees, particularly those in the upper percentiles.

IRT Based Equating Methods

IRT based equating methods is widely used among test practictioner based on the plusses derived from it. The procudure in IRT equating method is considered not too complex since IRT item and ability parameters are typically estimated separately for the two test forms, resulting in two different ability scales. However, in order to perform IRT based equating. Parameters must be on the same scale.

The Mean/Sigma method, which adjusts scores based on the means and standard deviations of two test forms, has been found to be effective in situations where the test forms are similar in difficulty. However, its effectiveness decreases when there are significant differences in the distributions of the two test forms. A study by Dilek, Atalay Kabasakal, and Gören (2025) examined the performance of the Mean/Sigma method in large-scale testing scenarios. They

found that while the method produced reliable equating results in cases where the test forms were of similar difficulty, it was less accurate when the difficulty level differed significantly between the test forms. For example, in an educational testing program, the Mean/Sigma method resulted in small equating errors (less than 1 scale point) when the two test forms were administered to a similar population of students. However, when the difficulty of the two forms varied, the equating errors increased.

The Mean-Mean Method has been widely used in high-stakes testing environments, particularly when there is an assumption that test forms have equal difficulty. Empirical results indicate that the method performs well when test forms are comparable, but can be less accurate when there is significant variation in test difficulty or item characteristics. An empirical study by Kolen and Brennan (2014) evaluated the Mean-Mean method in the context of a state-wide academic assessment. They observed that the method yielded accurate results when the test forms were designed to be equivalent in difficulty. For example, they found that equating errors were minimal (less than 0.5 standard deviation units) when the tests were aligned in difficulty. However, when test forms differed in difficulty, the equating procedure introduced substantial errors, which affected the test scores of higher-achieving students.

The Haebara Method, based on item response theory (IRT) and designed for nonlinear equating, has been shown to provide more accurate results than linear methods, especially in cases where the test forms differ in difficulty. Empirical studies demonstrate that this method is particularly effective in adjusting for varying item characteristics. Haebara (1980) demonstrated the effectiveness of his equating method by applying it to a set of reading comprehension tests. The study compared the equating outcomes between the Haebara method and simpler methods (e.g., linear equating). Results indicated that the Haebara method produced more accurate and reliable equating results when there were significant differences in test item characteristics. For instance, when the reading comprehension test was administered to different groups of students, the Haebara method was able to produce adjusted scores that were consistent across various levels of student ability, outperforming the Mean/Sigma method.

The Stocking-Lord Method has been shown to effectively account for differences in item characteristics and is frequently applied when there is no overlap between test items. This method has been shown to perform well in both standardized testing environments and in contexts with varying levels of item difficulty. Stocking and Lord (1983) conducted a study comparing the performance of the Stocking-Lord method with the Mean/Sigma and Haebara methods. Their study showed that the Stocking-Lord method provided more accurate equating results when item difficulties varied significantly between two test forms. Specifically, in a study involving a national certification exam, the Stocking-Lord method was able to adjust for item characteristic discrepancies and resulted in a more accurate alignment of scores across different test versions. The equating errors were found to be less than 0.2 standard deviation units, significantly lower than those observed with simpler equating methods.

Requirements for Test Equating In Educational Assessment

Holland and Dorans, (2006) point out that five requirements are widely viewed as necessary for an equating requirements which are:

1. The Equal Construct Requirement:The two tests should both be measures of the same construct (latent trait, skill, ability). For this condition to be met, the two tests need to be parallel and equal.

2. The Equal Reliability Requirement:The two tests should have the same level of reliability.
3. The Symmetry Requirement:The equating transformation for mapping the scores of Y to those of X should be the inverse of the equating transformation for mapping the scores of X to those of Y.
4. The Equity Requirement: It should be a matter of indifference to an examinee as to which of two tests the examinee actually takes.
5. The Population Invariance Requirement:The equating function used to link the scores of X and Y should be the same regardless of the choice of (sub) population from which it is derived.

Neil Dorans,Tim Moses and Daniel Eignor (2012) affrimed that with respect to best practices, requirements 1 and 2 mean that the tests need to be built to the same specifications, while requirement 3 precludes regression methods from being a form of test equating. Lord (1980) argue that requirement 4 implies both requirements 1 and 2. Requirement 4 is, however, hard to evaluate empirically and its use is primarily theoretical (Hanson, 1991; Lord, 1980). As noted by Holland and Dorans (2006), requirement 5, which is easy to assess in practice, also can be used to explain why requirements 1 and 2 are needed. If two tests measure different things or are not equally reliable, then the standard linking methods will not produce results that are invariant across certain subpopulations of examinees. Dorans and Holland (2000) used requirement 5, rather than requirement 4, to develop quantitative measures of equatability that indicate the degree to which equating functions depend on the subpopulations used to estimate them. For example, a conversion table relating scores on a mathematics test to scores on a verbal test developed on data for men would be very different from one developed from data on women, since, women tend to do less well than men on mathematics tests as indicated by litrature.

Test equating design

Equating Design Methods

In common, a test equating process consists of two important components: an equating design and equating methods. Equating design refers to a plan to collect equating data. It is sometimes called data collection design. This design include: Single Group Design (SGD), The Counterbalancing design (CD), Equivalent Groups Design (EGD) and Anchor Test Design which is also called non-equavlient anchor test (NEAT). The equating methods are the approches that can be used in implementing test equating, they are: Clasical Test Theory (CTT) method and Item Response Theory (IRT) method.

Practical application of test equating

Test Equating is an essential tool in educational assessment due to the critical role it plays in several key areas: establishing validity across forms and years; fairness; test security; and, increasingly, continuity in programs that release items or require ongoing development. However, equating can be viewed as a technical procedure or process conducted to establish equivalent scores on parallel forms of test, allowing them to be used interchangeably. It is an important aspect of establishing and maintaining the technical quality of a testing program by directly impacting the validity of assessments the degree to which evidence and theory support the interpretations of test scores. When two test forms (A and B) have been successfully equated, educators can validly interpret performance on one test form as having the same substantive meaning compared to the equated score of the other test form

The basis for equating is straightforward. When two tests, Form A and Form B, have been administered to different sets of examinees and scores on Form A are higher than those on Form B, several factors could explain such a difference. It could be that the difficulty index of Form A is higher than that of Form B, or examinees who took Form A were more proficient than those who took Form B. The necessity for equating arises in such circumstances. Subsequently, equating statistically adjusts for difficulty scores obtained on different test forms so that they are equivalent. Recent studies have emphasized the importance of equating in ensuring fairness and comparability across different test forms. For instance, research by Wallin and Wiberg (2024) introduced propensity score methods for local test score equating, highlighting their effectiveness in adjusting for group differences when anchor tests are not available. Similarly, Sappl (2023) explored various approaches to vertical equating, demonstrating how simultaneous analyses can improve efficiency in linking tests across grades. These advancements underscore the critical role of equating in attributing differences in score distributions to actual ability differences among examinees..

This concern helps to determine whether one cohort performed higher or lower on the test than the other cohorts. Equating is also instrumental in maintaining examination standards across test forms. In most testing companies and examination bodies, standards on examinations are usually set by experts through a well-controlled process. Oftentimes, performance descriptors are used to characterize the behavior of a borderline examinee on a test and cut scores are usually set through such a characterization. In fact, this is the way cut scores are oftentimes given meaning by basing them on judgments about the adequacy of test performance (i.e., on performance levels). Unfortunately, it is usually impractical to set standards on every test form. Therefore, test versions are equated so that the ability level associated with a cutoff point set on one test form remains constant over the subsequent administrations.

Conclusion

In conclusion, test equating serves as a cornerstone in the development and administration of fair and valid assessments. Its applications extend beyond mere statistical adjustments, playing a pivotal role in ensuring that test scores accurately reflect examinees' abilities and are free from biases or inconsistencies. As testing environments evolve, the continued refinement and application of equating methods will be essential in upholding the standards of fairness and validity in assessments.

The way forward on application of test equating

Practical Applications of Test Equating

Test equating is a critical process in educational and psychological assessments, ensuring that scores from different test forms are comparable and fair. Its applications are particularly pertinent in high-stakes testing environments where decisions based on test scores can significantly impact individuals' educational and professional trajectories. Below are key areas where test equating plays a vital role:

1. Ensuring Score Comparability across Different Test Forms

Equating allows for the comparison of scores from different test forms, ensuring that scores are interchangeable and reflect the same underlying construct. This is particularly important when multiple versions of a test are administered over time or across different groups. Recent advancements have introduced methods such as propensity score-based equating, which adjust for group differences when anchor tests are unavailable, thereby enhancing fairness in score comparisons.

© IIJP 2025 | Volume 1, Issue 4, pp. 312-327 | ISSN: 2636-4484

2. Maintaining Score Consistency over Time

Over time, changes in test items or populations can lead to shifts in test difficulty, a phenomenon known as item parameter drift (IPD). Equating methods help detect and correct for these drifts, ensuring that scores remain consistent and valid across different administrations. Studies have shown that IPD can lead to biased estimates of ability, highlighting the importance of regular equating to maintain score integrity.

3. Supporting High-Stakes Decision Making

In high-stakes testing scenarios, such as college admissions or professional certifications, equating ensures that decisions are based on fair and comparable scores. This is crucial for maintaining the validity and fairness of decisions that can significantly affect individuals' futures. Research emphasizes the need for rigorous equating procedures to uphold the credibility of high-stakes assessments.

4. Detecting and Addressing Security Breaches

Test equating can also play a role in identifying and mitigating the effects of security breaches, such as item exposure or cheating. By monitoring for inconsistencies in score distributions and item performance, equating procedures can help detect anomalies that may indicate compromised test integrity. This proactive approach aids in maintaining the security and fairness of the testing process.

5. Enhancing Test Security through Item Generation

Advancements in technology, such as automated item generation (AIG), have facilitated the creation of multiple test forms with similar difficulty levels. Equating these forms ensures that scores are comparable, thereby enhancing test security by reducing the likelihood of cheating and score inflation. AIG, combined with equating, offers a robust approach to maintaining test integrity.

Refrences

Agah, J. J. (2013). *Relative Efficiency of Test Scores Equating Methods in the Comparison of Students Continuous Assessment Measures* (Doctoral dissertation, UNN).

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (4th ed.). American Educational Research Association.

Angoff, W. H. (1971). *Scales, norms, and equivalent scores*. Educational Testing Service.

Aşiret, M. (2023). Impact of missing data on chain equating methods. *Educational Measurement Journal, 45*(2), 112–130.

Babcock, B., Albano, A., & Raymond, M. (2012). Equipercentile equating in large-scale assessments. *Assessment & Evaluation in Higher Education, 37*(5), 567–581.

Brauneis, R., & Goodman, E. P. (2018). Transparency and open testing. *Harvard Journal of Law & Technology, 31*(2), 567–610.

Dilek, A., Atalay Kabasakal, H., & Gören, M. (2025). Performance of the Mean/Sigma method in large-scale testing. *Journal of Educational Measurement, 62*(1), 34–52.

Dorans, N. J., & Holland, P. W. (2000). Quantifying equatability: Population invariance in score conversions. *Journal of Educational Measurement, 37*(1), 1–15.

Dorans, N. J. (2004). Linking scores: Methodologies and issues. *Applied Measurement in Education, 17*(2), 123–156.

Curcin, M., & Lee, M. W. (2025). Evaluating accuracy and bias of different comparative judgment equating methods against traditional statistical equating. *Frontiers in Education, 10,* Article 1538486. https://doi.org/10.3389/feduc.2025.1538486

González, J., & Wiberg, M. (2017). Test equating methodologies in educational assessment. *Psychometrika, 82*(3), 657–679.

Gübeşi, S., & Uyar, A. (2020). Equipercentile equating in skewed score distributions. *Educational Assessment, 25*(3), 198–212.

Haebara, T. (1980). Equating test scores based on item response theory. *Japanese Psychological Research, 22*(3), 127–144.

Hanson, B. (1991). Requirements for equating: A theoretical perspective. *Journal of Educational Measurement, 28*(1), 1–12.

Holland, P. W., & Dorans, N. J. (2006). *Linking and equating test scores*. Springer.

Issayeva, T. (2021). Test equating for evolving educational standards. *Assessment in Education: Principles, Policy & Practice, 28*(5), 621–639.

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

Leôncio, M., Smith, J., & Zhang, L. (2022). Classical and IRT-based equating methods: A comparative study. *Psychometrika, 87*(2), 456–480.

Livingston, S. A. (2004). Linear and equipercentile equating in high-stakes testing. *Journal of Educational Measurement, 41*(1), 1–15.

Livingston, S. A., & Kim, S. (2009). Linear equating revisited. *Applied Measurement in Education, 22*(3), 265–282.

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

Neil Dorans, T. M., & Eignor, D. R. (2012). Best practices in test equating. *Educational Measurement: Issues and Practice, 31*(4), 18–26.

Ryan, K., & Brockmann, R. (2009). Open testing and score comparability. *International Journal of Testing, 9*(2), 101–120.

Sappl, T. (2023). Advancements in vertical equating across grades. *Educational Measurement Journal, 63*(2), 112–135.

Sireci, S. G. (2005). The role of test equating in assessment fairness. *Journal of Educational Measurement, 42*(4), 271–280.

Stocking, M. L., & Lord, F. M. (1983). Developing a procedure for equating tests by IRT. *Journal of Educational Measurement, 20*(2), 69–78.

Sun, H., & Wang, Y. (2023). Equipercentile equating under non-normal distributions. *Educational Assessment, 28*(4), 345–367.

Sun, T., & Kim, S. Y. (2024). Evaluating equating methods for varying levels of form difference. *Educational and Psychological Measurement, 84*(3), 510–529. https://doi.org/10.1177/00131644231176989

Uzun, S. (2025). Differential item functioning in PISA 2018 mathematics: Implications for equating. *International Journal of Assessment in Education, 32*(1), 22–40.

Valentina, R., Marie, M., & Mariagiulia, S. (2017). Latent variable modeling and educational assessment. *Frontiers in Psychology, 8*, 1123. https://doi.org/10.3389/fpsyg.2017.01123

van Davier, M., Holland, P., & Thayer, D. (2004). *Item response theory for test equating*. Springer.

von Davier, M., et al. (2017). Equipercentile and chain equating in teacher certification exams. *Journal of Educational Measurement, 54*(2), 210–229.

von Davier, M., et al. (2024). Digital assessment and equating innovations. *Assessment in Education: Principles, Policy & Practice, 31*(3), 345–369.

Wallin, A., & Wiberg, M. (2024). Propensity score methods for local test score equating. *Psychometrika, 89*(1), 45–68.

Wyse, A. E. (2018). Rasch-based equating of Angoff ratings. *Journal of Educational Measurement, 55*(2), 234–250.