Asian Journal of Assessment in Teaching and Learning
*Vol 14, Issue 1, 2024 (70-84) /eISSN **2821-2916***

Test Equating in Educational Assessment: A Comprehensive Framework for Promoting Fairness, Validity, and Cross-Cultural Equity

Caroline Ochuko Alordiah*, John Oji
Faculty of Education, University of Delta, Agbor, Nigeria
*email: caroline.alordiah@unidel.edu.ng

Received: 13 April 2024; Accepted: 28 July 2024; Published: 30 June 2024

To cite this article (APA): Alordiah, C., & Oji, J. (2024). Test Equating in Educational Assessment: A Comprehensive Framework for Promoting Fairness, Validity, and Cross-Cultural Equity. *Asian Journal of Assessment in Teaching and Learning*, *14*(1), 70–84. https://doi.org/10.37134/ajatel.vol14.1.7.2024

To link to this article: https://doi.org/10.37134/ajatel.vol14.1.7.2024

Abstract

This study presents a comprehensive conceptual framework for equating in educational assessment, aimed at enhancing the accuracy, validity, and fairness of equating outcomes. The framework emphasizes the importance of considering sample characteristics, statistical assumptions, model fit, advancements in equating methodology, the integration of technology, and the factors of equity and fairness. By incorporating these elements, educational institutions can improve their equating practices and support equitable and fair evaluation processes. The framework also impacts policy-making and educational assessment procedures, providing a foundation for evidence-based policies that promote accountability and effective evaluation. Policymakers can use this framework to develop policies that ensure fair and valid assessment practices. Additionally, the study highlights the critical role of empirical research in validating and refining the framework, advocating for the exploration of cross-cultural equating methodologies to address diverse cultural contexts in education. To further advance the profession, the study suggests conducting empirical studies, embracing technology, fostering collaboration, increasing reporting standards, training practitioners, and monitoring equating practices. These efforts will help ensure more accurate, fair, and valid equating outcomes. This study offers valuable insights into equating in educational assessments, providing a robust basis for enhancing fairness, validity, and cross-cultural equity in educational evaluations.

Keywords: Test Equating, Educational assessment, Fairness, Validity, Cross-cultural equity

INTRODUCTION

Test equating ensures fairness and comparability in the assessment of student achievement, which is a crucial aspect of educational assessment. Evaluation of student performance, guidance in making decisions about education, and the success of educational initiatives all depend on the use of educational evaluations. Test equating makes it possible to compare people, groups, and times in a meaningful way and to interpret test results accurately (Huggins, 2014). The understanding of test equating, a crucial aspect of educational evaluation, lacks depth in theoretical exploration despite its paramount importance. Current literature predominantly focuses on practical methods for equating different test formats, neglecting thorough examination of the theoretical foundations behind these strategies (Leôncio et al., 2022; Born et al., 2019; Yuan et al., 2011). This gap impedes the development of a comprehensive understanding of test equating and its implications for educational assessment (Sansivieri et al., 2017). Further research is needed to elucidate the theoretical frameworks underlying equating methodologies, thereby enhancing the validity and reliability of assessment practices. By delving into the theoretical underpinnings, researchers can uncover insights that contribute to more effective and equitable evaluation methods. Consequently, addressing this knowledge gap is imperative for advancing educational assessment and ensuring fair and accurate evaluation outcomes.

In order to close this research gap, a solid conceptual framework for test equating is being built in this paper. The framework tackles the difficulties involved in equating, integrates and synthesises several equating approaches, and looks into new problems in the field. This framework will provide a clearer understanding of the complexity involved in test equating and direct future study and practice in the subject by creating a thorough theoretical foundation. Using well-known theories like Multidimensional Item Response Theory (MIRT), Item Response Theory (IRT), and Classical Test Theory (CTT), the conceptual framework enables a methodical investigation of the theoretical foundations of test equating. The framework aims to clarify the fundamental assumptions, principles, and limitations of various equating approaches by providing an in-depth explanation of their theoretical underpinnings. Moreover, the conceptual framework includes a thorough comprehension of the equating procedure, encompassing popular methods like Test Score Linking Methods, Anchor Test Design, and Common-Item Nonequivalent Groups Design (CING). Every method was analysed in light of the conceptual framework, enabling a thorough evaluation of its advantages, disadvantages, and suitability for various testing situations.

The difficulties with test equating were also included in the conceptual framework. These difficulties included the effects of sample variables (such as size and heterogeneity) and test form features (such as item overlap, test length, and differential item functioning) on equating processes. This research attempts to offer insights into the possible sources of bias and mistake in equating and suggest techniques for limiting their consequences by addressing these problems within the framework. In order to guarantee the validity and fairness of educational assessments, the framework also emphasised how crucial it is to address concerns of equity, fairness, and cultural considerations in test equating.

The goals of this study are to improve the validity and fairness of educational evaluations, advance the field, and support evidence-based methods. By means of a methodical analysis of the theoretical and conceptual underpinnings, this paper seeks to furnish scholars, professionals, and legislators with an invaluable tool for comprehending, executing, and refining test equating protocols. This study uniquely constructs a comprehensive conceptual framework for test equating by integrating both theoretical and practical dimensions, addressing a gap in existing research. Unlike prior work focused primarily on method application, it delves into the theoretical foundations of equating methods using Multidimensional Item Response Theory (MIRT), Item Response Theory (IRT), and Classical Test Theory (CTT). It also evaluates practical procedures like Test Score Linking Methods, Anchor Test Design, and Common-Item Nonequivalent Groups Design (CING), identifying potential biases and errors. Moreover, the study emphasizes equity, fairness, and cultural sensitivity in test equating, highlighting the need for culturally fair assessments. This dual focus on theory and practice, combined with an emphasis on fairness, sets this research apart, offering a robust foundation for advancing educational assessment and promoting more effective and equitable evaluation methods.

THEORETICAL FOUNDATION OF TEST EQUATING

Classical Test Theory

One of the main theoretical pillars in the subject of test equating is Classical Test Theory (CTT). It includes fundamental ideas and presumptions that direct the process of equating. The true score plus the error score are added to determine the observed test score in CTT. The test performance's random variations are captured by the error score, whereas the true score reflects the underlying skill or attribute being tested (Austin, 2019).

The real score and error score are assumed by CTT to be independent and regularly distributed. Additionally, it makes the assumption that the measurement error is constant throughout the whole true score range. These presumptions serve as the foundation for equating techniques, which seek to prove the equivalentity of various test formats or administrations (Lakens et al., 2018; Alordiah, 2015). But when it comes to test equating, CTT has come under fire. Its dependence on the presumption of a linear relationship between the observed and real scores—which might not hold true in all assessment contexts—is one of the main objections. Furthermore, CTT does not specifically take into account the features of individual test items or how differently they work across several groups or administrations (Algina, 2015).

In order to overcome these drawbacks, scholars have created more adaptable and sophisticated equating models using alternative theoretical frameworks as Multidimensional Item Response Theory (MIRT) and Item Response Theory (IRT). These models consider the characteristics of individual items on the test, the range of difficulty levels, and the possibility of multidimensionality in the construct being measured (Foster, 2019). Although CTT has been utilised extensively and offers a strong basis for comprehending test results, its shortcomings when it comes to test equating call for the investigation of other theories and methods. The criticism of CTT emphasises how crucial it is to take into account more intricate and thorough frameworks that can better reflect the intricacies of comparable processes and improve the precision and equity of educational assessments (Algina, 2015).

Item Response Theory (IRT)

A different theoretical framework for test equating that overcomes some of the drawbacks of Classical Test Theory (CTT) is provided by Item Response Theory (IRT). Understanding the relationship between a person's answer to a test item and the underlying ability or characteristic being tested is made easier with the help of IRT, which offers a more flexible and nuanced framework (Alordiah, 2022; Brzezińska, 2018). IRT is predicated on a number of important ideas and presumptions. One of the basic tenets is that the likelihood of answering an item correctly depends on both the person's aptitude and its features. IRT models evaluate an examinee's ability level based on their answer pattern, accounting for the qualities of particular test items, such as their difficulty and discriminatory power (Mair, 2018; Alordiah, 2015).

Unlike CTT, which assumes a linear relationship between observed and true scores, IRT models allow for more precise estimation of examinee abilities along a continuum. This is achieved through the use of item response curves, which provide information about the probability of a correct response at different levels of the underlying trait (Pena et al, 2018). IRT's capacity to take item attributes and their influence on the equating process into consideration is one of its main benefits when it comes to test equating. IRT-based equating techniques can account for the differentiating power and difficulty levels of items in various test forms and administrations. This makes it possible to compare test results more fairly and accurately, especially in cases where there may be differences in the format or item content of the test forms.

IRT models also enable the analysis of item attributes such item bias or differential item functioning, which contributes to a more thorough knowledge of test equating. These investigations can assist in locating possible causes of measurement bias and guarantee the impartiality and fairness of equating processes (Glas, 2014). IRT is applicable to test equating in a way that goes beyond conventional equating techniques. Test characteristic curve (TCC) equating and item characteristic curve (ICC) equating are two examples of IRT-based equating techniques that provide more adaptable options for equating tests with various item structures or answer formats (Leôncio et al, 2022).

A strong theoretical framework for test equating that overcomes some of the drawbacks of classical test theory is offered by item response theory. IRT is a useful framework for improving the equating of tests in educational assessments because of its capacity to model the relationship between item characteristics and examinee abilities, as well as its adaptability to various test formats and handling of measurement biases (Hori et al., 2020).

Multidimensional Item Response Theory (MIRT)

A theoretical framework called Multidimensional Item Response Theory (MIRT) expands on Item Response Theory (IRT) by taking into account the existence of several dimensions or latent qualities that are being measured in an exam. MIRT is aware that a lot of tests try to gauge intricate concepts with several facets, including the ability to solve mathematical problems using both procedural knowledge and problem-solving techniques. Since MIRT offers a more complex and all-encompassing method of comprehending and simulating the link between item answers and latent features, it is especially pertinent to test equating. MIRT provides additional flexibility in equating tests with varying item structures or content coverage and enables a more realistic depiction of the underlying construct being tested by taking into account several dimensions (Kim, 2022).

When equating methods, using MIRT may have a number of advantages. First of all, MIRT makes it possible to estimate distinct latent trait scores for every dimension, offering a more thorough knowledge of a person's capabilities on a variety of dimensions. More accurate equating outcomes may arise from this improved measurement precision, especially when equating tests with different dimensional coverage. Furthermore, MIRT can assist in revealing significant connections between the dimensions under study, enabling a more thorough examination of the construct's underlying structure. This can help with equating choices and guarantee that the multidimensionality of the concept is taken into account during the equating process (Kim, 2022).

Nevertheless, there are several difficulties when using MIRT to equate operations. The estimation of parameters in multidimensional models, which might be more computationally intensive than in unidimensional models, is one of the primary problems. Because of the modelling process's increasing complexity, precise parameter estimates, and trustworthy equating outcomes must be ensured (Lee, 2013). Furthermore, compared to unidimensional equating, the interpretation of data from multidimensional equating could be more complex. It can be more difficult to comprehend and communicate the meaning of equating outcomes in numerous dimensions, particularly when those dimensions have differing relative importance or relevance. Notwithstanding these difficulties, using MIRT to equate procedures has more advantages than disadvantages. MIRT offers a more accurate and thorough depiction of abilities by taking into consideration the multidimensionality of the construct being tested, which improves equating outcomes. By guaranteeing the fairness and comparability of test results across several dimensions, MIRT can help to improve equating procedures when parameter estimate is done with great care and interpretation is clear (Kim et al., 2020).

CONCEPTUAL FRAMEWORK FOR TEST EQUATING

Development of a comprehensive conceptual framework

To direct the study process and comprehend the intricate relationships and mechanisms involved, a thorough conceptual framework for test equating must be developed. This conceptual framework is presented in Table 1. This framework functions as a road map, outlining the important factors, connections, and fundamental mechanisms involved in the process of equating. A comprehensive analysis of the body of research on test equating, theoretical stances, and empirical data are all considered in the building of this framework. Researchers can create a conceptual framework that offers a thorough and organised understanding of the equating process by combining and synthesising information from these many sources.

Identification and classification of key variables in the best equating process

It is crucial to recognise and categorise the important factors that are crucial to the test equating procedure within the conceptual framework. These variables may consist of a variety of elements, including test formats, item attributes, examinee traits, and statistical techniques applied during equating. The various components involved in test equating are arranged and classified to some extent by the classification of these variables. Test forms can be divided into groups according to factors including response formats, difficulty levels, and subject coverage. It is possible to categorise item properties according on their type, discrimination, or difficulty. Examinee traits may include things like motivation, test-taking prowess, or past knowledge.

Construction of relationships and connections between variables within the framework

The next stage is to build the linkages and relationships between the important variables inside the conceptual framework after they have been recognised and categorised. This entails being aware of the interactions and influences that these factors have on one another while equating. To comprehend how various item attributes affect the equating process, for instance, the relationship between test forms and item properties might be investigated. In a similar vein, one might investigate the link between examinee features and equating approaches to find out how individual differences influence the selection and efficacy of equating methods. Through the establishment of these connections, the conceptual framework offers a comprehensive perspective on the equating procedure and assists in recognising the different elements that must be taken into account when carrying out equating studies or putting equating procedures into action.

Explanation of the underlying mechanisms and interactions within the conceptual framework

The last component of the conceptual framework is an explanation of the underlying mechanisms and variable interactions. This entails being aware of the psychometric concepts, statistical models, and theoretical underpinnings that underpin the equating procedure. The conceptual framework, for instance, describes how the theoretical underpinnings of equating techniques are provided by Classical Test Theory (CTT) or Item Response Theory (IRT). It also clarifies the application of statistical models to build links between test forms, such as the anchor test method and the equipercentile approach. The paradigm also emphasises how factors interact, for example, how the choice of equating method affects the accuracy and fairness of the equating outputs or how the anchor item selection affects the equating process. The conceptual framework offers a greater knowledge of the equating process by elucidating these underlying mechanisms and relationships. It also assists researchers in conducting studies, analysing data, and making well-informed judgements in the field of test equating. Creating a thorough conceptual framework for test equating is crucial to encouraging an organised and methodical approach to the field's practice and research. It facilitates the equating process' complexity, identifies important factors, forges connections between them, and explains the underlying mechanisms that underpin the equating results for academics and practitioners (Campbell, 2019).

The conceptual framework for test equating includes several important variable categories (Table 1). The first category is "Test Forms," which considers factors such as topic coverage, difficulty levels, and response styles. Test forms play a crucial role in determining item attributes and equating results. The next category is "Item Properties," which includes item kind, difficulty, and discrimination. These variables significantly impact examinee performance, equating outcomes, and test form selection. "Examinee Characteristics" is another category that encompasses variables like motivation, test-taking techniques, and past knowledge. These characteristics have a significant impact on item performance and equating results.

The "Equating Methods" category includes important techniques such as the Equipercentile Method, Anchor Test Method, Item Response Theory (IRT), and Test Characteristic Curve (TCC) equating. These methods help ensure comparability between different test forms, considering multidimensionality and item properties. The "Theoretical Foundations" category includes item response theory (IRT) and classical test theory (CTT), which guide the interpretation of equating results. Understanding the assumptions and constraints of these frameworks is crucial for a valid and correct equating process.

The "Statistical Models" category includes variables like item response theory (IRT), generalizability theory, and multidimensional item response theory (MIRT). These models are used to determine equating functions and estimate examinee and item parameters. The "Equating Design" category considers factors such as anchor test selection, matching criteria, and sample processes, which establish the foundation for equating and affect the results. Scaling procedures, including item calibration, connecting research design, and scaling techniques, fall under the "Scaling Procedures" category. These techniques ensure comparability of item parameters and create a connection between test forms. "Fairness Considerations" include variables like group comparisons, bias analysis, and differential item functioning (DIF), which address possible measurement bias and ensure equitable equating results. The "Validation Procedures" category includes item analysis, validity research, and reliability analysis to evaluate test form quality and interpret equated scores.

Lastly, the "Implementation Factors" category considers elements such as timelines, stakeholder involvement, and resources. These factors impact the feasibility and practicality of equating techniques. By considering these variables and their intricate relationships, researchers can conduct accurate and dependable test equating investigations that advance measuring and assessment research. It is crucial to prioritize test quality, fairness, validation, and practical limits to ensure accurate and equitable equating results.

Table 1. Conceptual framework for test equating

Variable Category	Key Variables	Relationships and Connections	Measurement Considerations	Implementation Considerations
Test Forms	Content Coverage, Difficulty Levels, Response Formats	Affect item characteristics and equating results	Make sure you sample items appropriately, reduce content overlap, and address potential construct underrepresentation	Consider practical constraints in test development, such as item availability and administration logistics
Item Properties	Discrimination, Difficulty, Item Type	Impact test form selection, equating outcomes, and examinee performance	Consider item quality and psychometric properties in form construction and equating decisions	Address item security concerns and potential item exposure across test forms
Examinee Characteristics	Prior Knowledge, Test-taking Skills, Motivation	Influence equating outcomes, impact item performance, and choice of equating methods	Think about the variety of examinee traits and how they affect the fairness equating process.	Address potential biases in equating outcomes related to examinee characteristics, such as gender or cultural background
Equating Methods	Equipercentile Method, Anchor Test Method, Item Response Theory (IRT), Test Characteristic Curve (TCC) equating	used to take multidimensionality into account, develop linkages between test formats, and account for item attributes	Consider the appropriateness of equating methods for different test characteristics and measurement objectives	Evaluate the computational requirements, technical expertise, and software availability for implementing different equating methods
Theoretical Foundations	Classical Test Theory (CTT), Item Response Theory (IRT)	Provide theoretical frameworks for equating methods and guide the interpretation of equating results	Consider the assumptions and limitations of the chosen theoretical framework	Ensure alignment between the chosen theoretical framework and the measurement goals of the equating process
Statistical Models	Item Response Theory (IRT), Generalizability Theory, Multidimensional Item Response Theory (MIRT)	Employed to estimate item and examinee parameters, account for multidimensionality, and establish equating functions	Consider the appropriateness and complexity of statistical models based on the nature of the test and equating objectives	Address potential challenges in parameter estimation, model fit, and assumptions underlying the chosen statistical models
Equating Design	Anchor Test Selection, Matching Criteria, Sampling Procedures	Determine the basis for equating, ensure the representativeness of test forms, and influence equating outcomes	Consider the representativeness and quality of anchor items, and the suitability of matching criteria	Address potential biases in the equating design, such as sample selection or nonresponse
Scaling Procedures	Item Calibration, Linking Study Design, Scaling Methods	Ensure the comparability of item parameters across test forms and establish the link between different forms	Consider the appropriateness of scaling procedures considering the test characteristics and equating goals	Address potential challenges in item calibration, linking study design, and the application of scaling methods

continued

Fairness Considerations	Differential Item Functioning (DIF), Bias Analysis, Group Comparisons	Address potential measurement bias, evaluate the fairness of equating outcomes across different groups, and inform equating decisions	Consider potential sources of bias and differential functioning across groups	Evaluate the fairness of equating outcomes for various examinee subgroups and address any identified biases
Validation Procedures	Item Analysis, Validity Studies, Reliability Analysis	Assess the quality of test forms, examine the validity and reliability of equating results, and inform the interpretation of equated scores	Consider the reliability and validity evidence supporting the equating process	Address potential limitations in the validation procedures and ensure the appropriateness of the chosen validation methods
Implementation Factors	Resources, Timelines, Stakeholder Involvement	Influence the feasibility and practicality of equating procedures, including the availability of resources, time constraints, and stakeholder perspectives	Consider the availability of resources, time constraints, and expertise required for implementing equating procedures	Engage stakeholders in the equating process and address their concerns and perspectives

APPROACHES TO TEST EQUATING

Common-Item No Equivalent Groups Design (CING)

A set of common items is included in both the old and new test forms as part of the CING design as presented in Table 2. By acting as a connecting element between the two forms, these shared elements enable the development of an equating relationship. The examinees' performance on the shared items in both forms is utilised to standardise the scores between the forms (Haberman, 2015). When there is a requirement to standardise scores between test versions or when the format or substance of the test is altered, the CING design comes in handy. It permits the estimate of equating functions, which guarantee score comparability by enabling the conversion of scores from one form to another (Gross et al., 2019).

Table 2. A step-by-step description of the CING Equating Process

Step	Description	Key Variables	Equating Methods	Statistical Models	Implementation Considerations
1	Test Form Construction and Selection	Construction of old and new test forms, Selection of common items	Content coverage, Difficulty levels, Item type	-	Consider relevance to construct, discrimination, and representativeness of items
2	Administration of Test Forms	Administer old and new test forms to different groups	-	-	Ensure similar administration conditions, minimize sources of measurement error
3	Scoring and Analysis	Score responses on common items, Obtain item scores or IRT parameters	Item properties (Discrimination, Difficulty), Item response theory (IRT)	-	Accurate scoring and estimation of item parameters
4	Equating Method Selection	Choose equating method based on test and equating goals	-	Equipercentile method, Linear regression, IRT-based methods	Consider equating goals, test characteristics, and available data

continued

Step	Description	Key Variables	Equating Methods	Statistical Models	Implementation Considerations
5	Estimation of Equating Relationship	Establish statistical relationship between performance on common items in both forms	-	Equipercentile method, Linear regression, IRT-based methods	Accurate estimation of equating relationship using appropriate statistical techniques
6	Score Conversion	Use equating relationship to convert scores from old form to new form scale	-	-	Accurate transformation of scores to ensure comparability and meaningful interpretation
7	Evaluation and Validation	Evaluate equating results, Conduct fairness analyses, Assess reliability and validity of equated scores	Fairness considerations, Validation procedures	-	Evaluate equating quality, fairness, reliability, and validity
8	Implementation and Reporting	Implement equating results, Report findings, Discuss limitations and evidence supporting validity and reliability	Implementation factors	-	Consider resources, timelines, stakeholder involvement, and clear reporting of equating process and results

In the test equating process in Table 2, researchers follow eight key steps. In Step 1, they design two test forms, ensuring common elements with good discrimination qualities and relevance to the construct being assessed. Step 2 involves administering the old and new forms to different sets of examinees, focusing on comparable administration circumstances. Step 3 involves scoring and analyzing the common items to obtain item scores or item response theory (IRT) parameters. In Step 4, researchers select an equating method based on test characteristics and equating goals. Step 5 involves establishing a statistical association between examinees' performance on the common items using the selected equating method. In Step 6, researchers convert scores from the old form to the new form's scale using the equating relationship established in the previous step. Step 7 focuses on evaluating the validity, reliability, and fairness of the equated scores, ensuring the accuracy and suitability of the equating process. Finally, in Step 8, researchers implement the equating results and report their findings, considering factors such as time, resources, stakeholder involvement, and transparent reporting. Throughout the process, researchers consider important variables such as content coverage, difficulty levels, item type, item characteristics, equating method selection, validation processes, fairness assessments, and implementation considerations. The goal is to ensure accurate and consistent equating results that can be applied effectively in measuring and assessment research.

Anchor Test Design

The anchor test is regarded as a "common reference" since all test formats employ the same elements. Researchers can create an equating relationship between the results on the various test forms and the anchor test scores by incorporating the anchor test into the equating process. This relationship makes it possible to convert scores between forms, which serves as a foundation for meaningful score interpretation and comparability. An essential component of the conceptual framework for equating is the anchor test design. The following actions can be taken to integrate it into the framework. This is clearly captured in Table 3.

In the process of test equating in Table 3, researchers follow eight key steps. Step 1 involves constructing test forms and an anchor test using common items, considering factors such as difficulty levels, discrimination, and relevance to the construct being tested. Step 2 focuses on administering the test forms and anchor test to different examinee groups under comparable circumstances. Step 3 entails scoring the anchor test items and analyzing item properties, such as discrimination and difficulty. In Step 4, researchers select an equating method, such as the equipercentile method or linear regression, to link the anchor test scores to the test form scores. Step 5 involves establishing a statistical correlation between the anchor test and test form scores based on the chosen equating method. Step 6 requires accurately converting anchor test scores to the scale of the test forms using the equating relationship. Step 7 involves evaluating the validity, reliability, and fairness of the equated scores, considering important variables such as validation processes and fairness assessments. Researchers must assess the quality, dependability, and validity of the equating procedure. Finally, in Step 8, researchers implement the equating results, taking into account factors like time, resources, stakeholder involvement, and transparent reporting. They provide evidence of the accuracy and consistency of the equated scores, while acknowledging the constraints of the equating process. Throughout the process, researchers prioritize important variables such as item properties, equating method selection, validation processes, fairness considerations, and implementation factors. The aim is to ensure accurate and reliable equating results that can be effectively applied in measuring and assessment research.

Table 3. Step-by-step guide on how to integrate Anchor Test Design on the conceptual framework

Step	Description	Key Variables	Equating Methods	Statistical Models	Implementation Considerations
1	Test Form Construction and Selection	Construct test forms and develop anchor test with common items	Content relevance, Discrimination, Difficulty	-	Consider construct coverage, item quality, and representativeness
2	Administration of Test Forms and Anchor Test	Administer test forms and anchor test to different groups	-	-	Ensure similar administration conditions for both test forms and anchor test
3	Scoring and Analysis	Score responses on anchor test items, Obtain anchor test scores or IRT parameters	Item properties (Discrimination, Difficulty), Item response theory (IRT)	-	Accurate scoring and estimation of item parameters for the anchor test
4	Equating Method Selection	Choose equating method for linking anchor test scores to test form scores	-	Equipercentile method, Linear regression, IRT-based methods	Consider equating goals, test characteristics, and available data
5	Estimation of Equating Relationship	Establish statistical relationship between anchor test scores and test form scores	-	Equipercentile method, Linear regression, IRT-based methods	Accurate estimation of equating relationship using appropriate statistical techniques
6	Score Conversion	Use equating relationship to convert anchor test scores to the scale of test forms	-	-	Accurate transformation of anchor test scores to ensure comparability and meaningful interpretation
7	Evaluation and Validation	Evaluate equating results, Conduct fairness analyses, Assess reliability and validity of equated scores	Fairness considerations, Validation procedures	-	Evaluate equating quality, fairness, reliability, and validity

continued

8	Implementation and Reporting	Implement equating results, Report findings, Discuss limitations and evidence supporting validity and reliability	Implementation factors	-	Consider resources, timelines, stakeholder involvement, and clear reporting of equating process and results

TEST SCORE LINKING METHODS

Methods for creating a connection between test results from various forms are known as test score linkage strategies. These techniques are pertinent to the conceptual framework of equating because they allow researchers to guarantee comparability across various test forms and equate test scores. Here is a summary of a few popular linking techniques:

Equipercentile Method

Finding equivalent percentiles on two or more test forms and connecting them is known as the equipercentile approach. The fact that this approach offers a simple means of establishing an equating relationship based on percentile ranks makes it pertinent to the conceptual framework. But it makes the assumption that the distribution of scores is the same for all test forms (Varas et al., 2020). The equipercentile approach is not too difficult to use and comprehend. It offers a straightforward method of connecting test results based on percentile ranks. When the distribution of scores is consistent among test formats, it may function effectively. The equipercentile approach makes the unavoidable assumption that test forms' score distributions are comparable. It may not take into account variations in item difficulty or examinee skill, nor does it take into account the underlying relationship between the scores (Sun, 2021).

Linear Regression

Fitting a regression equation that explains the link between the scores on the anchor test and the test forms being equal is known as linear regression. This approach is pertinent because it takes into consideration the variation in scores, enabling a more accurate calculation of the equating connection. But it makes the assumption that the anchor test and the test forms have a linear connection (Casson, 2014). Because linear regression takes score variability into account, it provides a more accurate estimate of the equating connection. Polynomial regression is a useful tool for capturing non-linear correlations. Additionally, it can shed light on how the anchor test and the test forms relate to one another (Albano, 2015). Linear regression assumes a linear relationship between the anchor test and the test forms, which may not always hold true. It relies on the assumption of linearity and homoscedasticity. It may be sensitive to outliers and may require a sufficient sample size to obtain stable estimates.

Item Response Theory (IRT)-Based Methods

IRT-based methods can account for variations in item characteristics and examinee abilities across test forms because they use models that explain the relationship between item responses and latent traits. Examples of these models are the mean-sigma method and the concurrent calibration method, which are relevant to the conceptual framework because they provide a flexible framework for equating (Brzezińska, 2016). Because IRT-based approaches take examinee abilities and item attributes into account across test forms, they offer a flexible framework for equating. They can take into consideration variations in guessing parameters, discriminating, and item complexity. They make it possible to estimate an IRT linking function, which offers a more accurate equating connection (Adetutu, 2022). A solid grasp of IRT models and assumptions is necessary to effectively apply IRT-based approaches. To get reliable parameter estimations, they might need a big sample size. They may need a lot of time and complicated computations. They rely on the presumption that test forms' item parameters are invariant (Reyhanlioğlu, 2020).

CHALLENGES IN TEST EQUATING

Test Form Characteristics

The length of a test form, which refers to the number of items included, can impact the equating process. Equipercentile procedures are sensitive to test duration, with longer test forms generally producing more consistent equating results. However, the impact of test length is less significant in linear regression and Item Response Theory (IRT) techniques (Leontaridou et al., 2019). Item overlap, which represents shared items across different test forms, is important for accurate equating. A sufficient number of common items strengthens the equating link. However, excessive item overlap can compromise the ability of a test form to measure different constructs. Item characteristics, such as difficulty, discrimination, and guessing parameters, can also affect equating. Differences in item characteristics between test forms can introduce measurement bias and reduce equating precision. It is essential to ensure that item characteristics are comparable across test forms for valid equating (Qiu et al., 2018).

To validate equating, it is crucial to establish equivalence between test forms. Variations in test length, item overlap, and item attributes can introduce measurement bias and undermine the equating process. Careful planning and construction of test forms, including the selection and development of items with comparable qualities, are necessary to reduce measurement error and improve equating accuracy. Additionally, considering how item attributes impact equating can help identify potential sources of error (Bais et al., 2019). Sufficient sample size is also important for reliable equating. Small sample sizes, especially when using IRT-based techniques, can lead to unstable equating estimates. Researchers should ensure an adequate sample size to produce trustworthy equating results. When selecting an equating method, the characteristics of the test forms and the equating objectives should be taken into account. Different equating techniques may vary in their sensitivity to test duration, item overlap, and item attributes. Researchers should carefully choose an appropriate equating approach based on these considerations (Wang et al., 2013). Furthermore, test form characteristics, particularly item overlap, can impact equating fairness. It is essential to prevent any biases that may penalize specific groups of examinees. Thorough analysis and fairness assessments should be conducted to address any potential biases resulting from test form characteristics.

Sample Characteristics

The required sample size for equating purposes varies depending on the equating method employed. Equipercentile approaches typically require larger sample sizes to ensure consistent equating results, especially when dealing with test forms that have different score distributions. IRT-based and linear regression techniques may require smaller sample sizes but can still impact the accuracy of the equations (Adhikari, 2021). The necessary sample size is also influenced by the desired level of equating precision. Higher precision requires larger sample sizes. Researchers must consider the trade-off between accuracy and real-world constraints, such as time and resource availability, when determining the acceptable level of inaccuracy in equating. The features of the test forms also affect the sample size requirements. Test forms with more items, higher complexity, or greater variability among examinees may necessitate larger sample sizes to obtain accurate equating estimates. Researchers should consider these unique characteristics when calculating the required sample size (Hajian-Tilaki, 2014).

Sample heterogeneity, associated with variations in examinee characteristics, such as age, gender, educational background, or language competence, should be addressed through subgroup studies. These studies help identify potential biases or differential item functioning (DIF) and improve equating accuracy. Failure to account for DIF can impact equating outcomes, as different subgroups with the same underlying ability may have differing odds of answering items successfully. Researchers should employ suitable DIF detection techniques and ensure fair equalization outcomes for all subgroups (Huggins, 2014). To enhance the external validity of equating, the sample utilized for equating should represent the population of examinees for whom the equating results will be generalized. A representative sample ensures that the equating results accurately reflect the performance of the target population. When sample heterogeneity is detected, statistical corrections such as propensity score matching and covariate adjustment can be employed to mitigate its effects and improve equating accuracy (Alba et al., 2016).

Statistical Assumptions and Model Fit

Equating models rely on several assumptions to accurately establish a connection between test forms and the underlying construct. These assumptions include the invariance of measuring attributes, such as item difficulty and discrimination, which are evaluated through measurement invariance analysis (Edwards et al., 2018). Linear equating models assume a constant relationship between scores on the anchor test and the test forms throughout the score range, while the assumption of homoscedasticity posits that score variability remains constant across different levels of the underlying concept (Wu, 2023). Equating models also often assume a normal distribution of test scores, which is crucial for precise parameter estimation and hypothesis testing (Qiu et al., 2019).

The overall fit of equating models to the data can be assessed using goodness-of-fit tests, such as chi-squared tests or model fit indices like RMSEA, CFI, and TLI, providing statistical evidence of how well the model represents the observed data (Qiu et al., 2019). Residual analysis, through methods like histograms and scatterplots, helps identify trends or anomalies that may indicate violations of assumptions or model misspecification (Schielzeth et al., 2020). Sensitivity analysis, which involves data simulations or deliberate introduction of assumption deviations, examines the stability and reliability of equating models when assumptions are broken, influencing model selection and understanding.

Diagnostic approaches, such as examining standardized residuals or using statistical methods like item fit analysis or differential item functioning identification, help identify items or subgroups that may defy assumptions, guiding focused model improvement or modification (Schielzeth et al., 2020). These assessment techniques collectively contribute to ensuring the accuracy and validity of equating models.

IMPLICATIONS OF THE CONCEPTUAL FRAMEWORK FOR EDUCATIONAL ASSESSMENT PRACTICES AND POLICY MAKING

The conceptual framework for equating procedures has wide-ranging implications beyond the study itself. Firstly, it serves as a comprehensive guide for educational institutions to ensure correct and consistent equating practices, leading to improved academic outcomes and more accurate test results. Secondly, it aids in making fair and reliable decisions regarding student performance, program placement, and cut-off scores, resulting in more meaningful educational policies and practices. Thirdly, the framework emphasizes the importance of incorporating justice, equity, and cultural factors in equating processes, promoting fairness and addressing disparities across student demographics. Fourthly, policymakers can utilize the framework to develop evidence-based assessment policies that align with best practices in equating, supporting accountability and educational reform. Lastly, the framework underscores the need for ongoing research and improvement in equating methodology, technology, and emerging concerns, ensuring that equating practices remain relevant and responsive to evolving evaluation requirements.

CONCLUSION

The conceptual framework presented in this study provides a comprehensive understanding of equating in educational assessment. It highlights the importance of sample characteristics, statistical assumptions, model fit, improvements in equating procedures, technological integration, and addressing justice, fairness, and cultural factors. By implementing these components, educational institutions can enhance the validity, fairness, and precision of equating results. Additionally, the framework has implications beyond research, informing educational assessment practices and policy development. Policymakers can create evidence-based policies that support accountability and effective assessment procedures. Continual research and framework improvement are crucial to maintain its relevance and effectiveness. Overall, the conceptual framework offers a strong foundation for implementing equating in educational evaluation and advancing equitable practices and policies.

RECOMMENDATIONS

While the conceptual framework provides a strong foundation, further empirical research is needed to validate and enhance it. This includes evaluating the effectiveness of different equating techniques, studying the impact of sample characteristics on equating results, and examining the real-world implications of incorporating equity and fairness considerations. Cross-cultural equating approaches should be explored to ensure validity and fairness across diverse cultural contexts. Comparative research can shed light on the effectiveness of equating techniques in different cultural settings. Technological advancements, such as data mining, artificial intelligence, and machine learning, should be investigated for their potential integration with equating to improve accuracy and efficiency. Collaboration among researchers, practitioners, and policymakers is crucial to address the complexities of equating. Establishing equating-focused forums, conferences, and research networks can facilitate knowledge exchange and interdisciplinary collaboration. Standardized reporting guidelines are essential to ensure transparency and reproducibility in equating research. Training programs should be funded to equip practitioners with the necessary skills for consistent and efficient equating procedures. Institutions should establish systems for ongoing evaluation and feedback to continuously improve equating methods.

REFERENCES

Adetutu, O., & Lawal, H. (2022). Applications of Item Response Theory models to assess item properties and students' abilities in dichotomous responses items. *Open Journal of Educational Development (ISSN: 2734-2050).* https://doi.org/10.52417/ojed.v3i1.304.

Adhikari, G. (2021). Calculating the Sample Size in Quantitative Studies. *Scholars' Journal*. https://doi.org/10.3126/scholars.v4i1.42458.

Alba, A., Alexander, P., Chang, J., Macisaac, J., DeFry, S., & Guyatt, G. (2016). High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes.. *Journal of clinical epidemiology*, 70, 129-35 . https://doi.org/10.1016/j.jclinepi.2015.09.005.

Albano, A. (2015). A General Linear Method for Equating with Small Samples.. *Journal of Educational Measurement*, 52, 55-69. https://doi.org/10.1111/JEDM.12062.

Algina, J., & Swaminathan, H. (2015). Psychometrics: Classical Test Theory. , 423-430. https://doi.org/10.1016/B978-0-08-097086-8.42070-2.

Alordiah, C. O. (2022). An examination of the latent constructs in a well-being scale for children: Application of Rasch Model. *University of Delta Journal of Contemporary Studies in Education*, *1* (2), 39-57.

Alordiah, C. (2015). Comparison of index of Differential Item functioning under the methods of Item Response theory and classical test theory in Mathematics. *An unpublished Ph. D thesis of Delta State University, Abraka, Delta State, Nigeria*.

Austin, J. (2019). Classical Test Theory and Music Testing. *The Oxford Handbook of Assessment Policy and Practice in Music Education, Volume 1*. https://doi.org/10.1093/OXFORDHB/9780190248093.013.21.

Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tóth, J., Douhou, S., Kieruj, N., Morren, M., & Vis, C. (2019). Can Survey Item Characteristics Relevant to Measurement Error Be Coded Reliably? A Case Study on 11 Dutch General Population Surveys. *Sociological Methods & Research*, 48, 263 - 295. https://doi.org/10.1177/0049124117729692.

Born, S., Fink, A., Spoden, C., & Frey, A. (2019). Evaluating Different Equating Setups in the Continuous Item Pool Calibration for Computerized Adaptive Testing. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01277

Brzezińska, J. (2018). Item response theory models in the measurement theory. *Communications in Statistics - Simulation and Computation*, 49, 3299 - 3313. https://doi.org/10.1080/03610918.2018.1546399.

Campbell, I. (2019). Test Equating Requirements from an SEM Perspective. *Multivariate Behavioral Research*, 54, 147 - 148. https://doi.org/10.1080/00273171.2018.1555748.

Casson, R., & Farmer, L. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clinical & Experimental Ophthalmology*, 42. https://doi.org/10.1111/ceo.12358.

Edwards, M., Houts, C., & Wirth, R. (2018). Measurement invariance, the lack thereof, and modeling change. *Quality of Life Research*, 27, 1735-1743. https://doi.org/10.1007/s11136-017-1673-7.

Foster, R. (2019). A generalized framework for classical test theory. *Journal of Mathematical Psychology*. https://doi.org/10.31234/osf.io/4j9vt.

Glas, C. (2014). Item response theory in educational assessment and evaluation., 31, 19-34. https://doi.org/10.7202/1025005AR.

Gross, A., Kueider-Paisley, A., Sullivan, C., & Schretlen, D. (2019). Comparison of Approaches for Equating Different Versions of the MMSE Administered in 22 Studies.. *American journal of epidemiology*. https://doi.org/10.1093/aje/kwz228.

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics*, 48, 193-204 . https://doi.org/10.1016/j.jbi.2014.02.013.

Haberman, S. (2015). Pseudo-Equivalent Groups and Linking. *Journal of Educational and Behavioral Statistics*, 40, 254 - 273. https://doi.org/10.3102/1076998615574772.

Hori, K., Fukuhara, H., & Yamada, T. (2020). Item response theory and its applications in educational measurement Part II: Theory and practices of test equating in item response theory. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14. https://doi.org/10.1002/wics.1543.

Huggins, A. (2014). The Effect of Differential Item Functioning in Anchor Items on Population Invariance of Equating. *Educational and Psychological Measurement*, 74, 627 - 658. https://doi.org/10.1177/0013164413506222

Kim, S., Lee, W., & Kolen, M. (2020). Simple-Structure Multidimensional Item Response Theory Equating for Multidimensional Tests. *Educational and Psychological Measurement*, 80, 125 - 91. https://doi.org/10.1177/0013164419854208.

Lakens, D., Scheel, A., & Isager, P. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1, 259 - 269. https://doi.org/10.1177/2515245918770963.

Lee, E. (2013). Equating multidimensional tests under a random groups design: A comparison of various equating procedures. . https://doi.org/10.17077/ETD.QPBFYMEI.

Leontaridou, M., Gabbert, S., & Landsiedel, R. (2019). The impact of precision uncertainty on predictive accuracy metrics of non-animal testing methods.. *ALTEX*. https://doi.org/10.14573/altex.1810111.

Leôncio, W., Wiberg, M., & Battauz, M. (2022). Evaluating Equating Transformations in IRT Observed-Score and Kernel Equating Methods. *Applied Psychological Measurement*, 47, 123 - 140. https://doi.org/10.1177/01466216221124087.

Mair, P. (2018). Item Response Theory. , 95-159. https://doi.org/10.1007/978-3-319-93177-7_4.

Pena, C., Costa, M., & Oliveira, R. (2018). A new item response theory model to adjust data allowing examinee choice. *PLoS ONE*, 13. https://doi.org/10.1371/journal.pone.0191600.

Reyhanlioğlu, Ç., & Doğan, N. (2020). An Analysis of Parameter Invariance according to Different Sample Sizes and Dimensions in Parametric and Nonparametric Item Response Theory. , 11, 98-112. https://doi.org/10.21031/epod.584977.

Schielzeth, H., Dingemanse, N., Nakagawa, S., Westneat, D., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N., Garamszegi, L., & Araya-Ajoy, Y. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11, 1141 - 1152. https://doi.org/10.1111/2041-210X.13434.

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A Review of Test Equating Methods with a Special Focus on IRT-Based Approaches. *Statistica*, *77*(4), 329–352. https://doi.org/10.6092/issn.1973-2201/7066

Sun, T., & Kim, S. (2021). Evaluating Six Approaches to Handling Zero-Frequency Scores under Equipercentile Equating. *Measurement: Interdisciplinary Research and Perspectives*, 19, 213 - 235. https://doi.org/10.1080/15366367.2020.1855034.

Varas, I., González, J., & Quintana, F. (2020). A Bayesian Nonparametric Latent Approach for Score Distributions in Test Equating. *Journal of Educational and Behavioral Statistics*, 45, 639 - 666. https://doi.org/10.3102/1076998620907381.

Qiu, Y., Liu, L., Lai, X., & Qiu, Y. (2019). An Online Test for Goodness-of-Fit in Logistic Regression Model. *IEEE Access*, 7, 107179-107187. https://doi.org/10.1109/ACCESS.2019.2927035.

Wang, L., Liu, Y., Wu, W., & Pu, X. (2013). Sequential LND sensitivity test for binary response data. *Journal of Applied Statistics*, 40, 2372 - 2384. https://doi.org/10.1080/02664763.2013.817546.

Wu, J., & Drton, M. (2023). Partial Homoscedasticity in Causal Discovery With Linear Models. *IEEE Journal on Selected Areas in Information Theory*, 4, 639-650. https://doi.org/10.1109/JSAIT.2023.3328476.

Yuan, S., Zhao, S., & He, Z. (2011). Test equating and model application. *2011 International Conference on Computer Science and Service System (CSSS)*, 3640-3643. https://doi.org/10.1109/CSSS.2011.5974614.