IJECA
International Journal of
Education & Curriculum Application
http://journal.ummat.ac.id/index.php/IJECA
ISSN 2614-3380 | Vol. 6, No. 3, December 2023

Equating Principles in Assessment: A Literature Review in the Context of Education and Assessment

Muh. Fitrah[1], Ilyas[2], Nur Rahmi Akbarini[3], Oscar[4], Edi Istiyono[5], Widihastuti[6]
[1,2,3,4,5,6]Graduate School, Research and Educational Evaluation, State University of Yogyakarta, Indonesia
muhfitrah.2023@student.uny.ac.id[1]

Keywords:
Equating principles; Educational measurement; Assessment analysis.

ABSTRACT

This literature review explores the fundamental concepts and methods related to equating principles in the context of educational assessment. Equating, as a complex statistical technique, plays an increasingly crucial role in ensuring fairness and accuracy in test-based assessments. The literature review methodology begins with the identification of key themes regarding equating principles in education and assessment. It utilizes keywords and scholarly databases to search for relevant sources, selecting those that are up-to-date and possess robust methodologies. The result is a deeper understanding of the concept of equating in the context of educational assessment. The study highlights the significant relevance of equating in education, where test results are often used for critical decision-making. Extensive discussions on the practical implications of implementing equating in educational policy and assessment, as well as its impact on students, teachers, and educational institutions, are presented. In conclusion, a critical understanding of equating is essential to ensure fair, consistent, and meaningful assessments in an ever-evolving educational landscape.

Article History:
Received: 01-10-2023
Revised : 06-11-2023
Accepted: 07-11-2023
Online : 01-12-2023

https://doi.org/10.31764/ijeca.v6i3.19381

A. INTRODUCTION

Assessment is a critical aspect of the education system (Wiliam & Thompson, 2017) and performance evaluation in various sectors (Bayo-Moriones et al., 2020). The assessment process is used to measure the achievements, skills, knowledge, and abilities of individuals in various contexts, such as schools, workplaces, or scientific research. The principles of assessment instruments refer to appropriate and valid methods (Scott et al., 2019; Johnson et al., 2021). Furthermore, good assessment instruments should be comprehensive, encompassing various relevant skills and knowledge related to the subject being assessed (Shofwanthoni et al., 2019; Blanco-Vogt & Schanze, 2014). The assessment instrument, as a principle of assessment, is an integral component of the education system and plays a crucial role in measuring students' achievements, skills, knowledge, and abilities (Wahhab & Rizko, 2019; Heinrichs-Graham et al., 2022). To ensure that such assessments are fair and reliable, fundamental principles in assessment need to be well-applied.

One of the principles that has emerged in the field of assessment is the principle of equating. This principle is closely related to efforts to compare assessment results across different tests or among various groups of individuals (Duarte & Rossier, 2008). The equating principle has become increasingly important in the context of globalization and high mobility, where comparisons and recognition of assessment results have significant implications (Papastephanou, 2005).

In the educational context, a fair and valid assessment approach is crucial in measuring students' progress from year to year, understanding the effectiveness of educational programs, and ensuring that all students have an equal opportunity to develop. At the professional level, employee performance assessment is also a key element in human resource management, recruitment, and career development. However, despite the importance of the equating principle in assessment, there are still challenges and issues to be addressed. These challenges include the development of accurate equating methods Kolen & Brennan (2004), the use of technology to support the assessment process, and policies that promote equality in assessment (Lucey et al., 2020). In addressing fairness issues in assessment, it is essential to consider various strategies that promote equality and provide opportunities for all students to demonstrate their learning outcomes (Nayir et al., 2019). One crucial aspect of fairness in assessment is the inclusion of various assessment strategies that take into account the diverse backgrounds and experiences of students.

Therefore, this paper will provide a deeper exploration of the equating principle in assessment, delving into the various methods and techniques used and their implications. Beyond that, this paper will also explore recent developments in the field of equating and the challenges that still exist in efforts to create fairer and more accurate assessments. It is hoped that this paper will offer valuable insights into the importance of the equating principle in assessment and how this principle can be effectively applied to support individual development and overall societal progress.

B. METHOD

The literature review method applied in this article begins with the identification of the main theme, namely, equating principles in the context of education and assessment. The initial step involves searching for relevant literature sources on this topic, including journal articles, books, research reports, and assessment guidelines. Various scholarly databases such as ERIC, ProQuest, and Google Scholar were used to search for literature related to equating principles. Keywords used in the search included equating principles, measurement, and assessment.

After gathering a sufficient number of literature sources, we conducted a selection based on inclusion criteria, prioritizing sources that were relevant, up-to-date, and had strong methodologies in the context of equating. We identified approaches and models that have been used in previous research to apply the equating principle in educational assessments. During the literature analysis, we paid attention to the main trends in the development of equating principles over time and and compared different approaches in terms of assessment application. The result of this literature review method is a deeper understanding of the equating principle in the context of education and assessment. We organized our literature review by summarizing key concepts, historical developments, and the application of equating principles in various educational assessment contexts.

C. RESULT AND DISCUSSION

1. Equating in Assessment

In various aspects of life, assessment plays a crucial role as a tool to measure performance, progress, and quality. Whether in the realm of education, employment, or other decision-making contexts, assessment serves as the foundation for making accurate and fact-based evaluations. Assessment is an essential component of education (Mardapi, 2017). Assessment is not merely about determining scores or numbers (Mau, 2020), rather, it is a complex process involving data collection, interpretation, and decision-making based on predefined objectives.

Asessment is the process of gathering information about individual or group behavior to evaluate the achievement of educational goals related to interesting educational variables (Popham, 2008). In the educational context, as stated by Erfianti et al., (2019), assessment is not limited to students but also includes the assessment of educators, teaching methods, and school administration (Singh et al., 2021; Gaytan & McEwen, 2007).

Furthermore, Putri & Istiyono (2017) explain that the concept of assessment in education is aimed at enhancing learning outcomes that align with the assessment objectives. Assessment for learning is used to monitor the knowledge that learners have acquired, taking into consideration learners' self-evaluation (Cassidy, 2007). Assessment as learning is employed to evaluate the achievement of learning goals from the beginning to the end of the learning process (Hung, 2019; Umar & Majeed, 2018).

In the assessment process, context is the key. In this regard, assessment is not merely about data collection; it also provides evidence of the achievement of specific objectives or standards used to make decisions about success or improvement. On the other hand, Immonen et al. (2019) assessment should be systematic, continuous, and integrated (Lyngsø et al., 2014). According to Gronlund (2003), assessment is the process of collecting, analyzing, and interpreting information to assess an individual's progress in achieving learning objectives (McDonald, 2002). In general, the interpretations of assessment by various experts indicate a process that involves collecting information, interpretation, and decision-making based on an individual's achievement of predefined goals, standards, or competencies.

Classroom learning encompasses the acquisition of knowledge, skill development, and shaping students' attitudes. High-quality learning refers to the improvement in the quality of student graduation. The quality of graduation, in turn, becomes a key indicator of educational quality. Assessment plays a central role in the education system because assessment results reflect the development or progress of education that can be measured over time. This allows for comparisons between different schools or regions. This process, known as equating in measurement terms, aims to standardize the level of educational achievement across various schools or regions.

The statistical process used to equate scores between two tests is known as equating. Equating is a statistical procedure used to understand the relationship between scores on two or more tests (Himelfarb, 2019). Kolen & Brennan (2004) explain that the equating process is carried out to adjust two or more tests that have equivalent content and difficulty levels. Using the same ability scale in test score equating has several benefits, including enabling test score evaluations, developing equatable tests, ensuring test security, and facilitating item bank development.

Humphry (2006) explains that in the procedure of test score equating, there are two main approaches: vertical equating and horizontal equating. Vertical equating is used to adjust test scores with different difficulty levels but measures the same type and content of skills. It is designed to measure students' skill level development or change over time. On the other hand,

horizontal equating is performed on parallel tests with similar content and difficulty levels and is administered to groups of students with equivalent skill levels. According to Kolen & Brennan (2004) to reduce inaccuracies in equating results, careful equating design is necessary. Several equating designs are available, such as single-group design, equivalent-group design, balanced-group design, and anchor design (Heri Retnawati, 2014). Each of these designs has its characteristics and advantages as well as disadvantages (Aminah, 2012).

The use of item response theory in test score equating is considered more representative than using classical test theory. Item response theory has the property of parameter invariance. This means that student ability parameters do not change depending on test parameters, and vice versa. Therefore, tests taken by students will remain on the same scale when the test information function is high. Item response theory has two models, namely the item response theory model for dichotomous data and the item response theory model for polytomous data. Dichotomous data only have two possible answers (correct or incorrect), while polytomous data have more than two possible answers (Embretson & Reise, 2000).

2. Principles of Equating in Assessment

In the book titled Item Response Theory and Its Application by Retnawati (2014) four fundamental equating principles, are explained. These principles include the principles of equality, population invariance, symmetry, and unidimensionality. These four principles of equating in assessment are critical principles that help ensure fair, objective, and non-discriminatory assessment practices.

a. Equality Principle

Ideally, equality in assessment should be considered at the earliest stages of planning, development, or implementation (Were et al., 2019). Equality in assessment refers to actions taken by educators to ensure that all students have equal opportunities (Elkhoury et al., 2023), and the assessment instruments and testing processes should align with the curriculum content, ensuring that they do not disadvantage students' learning opportunities.

The term “equating” differs from “equality.” Equating encompasses the entire statistical procedures used to understand the relationship between scores on two or more tests, while equality refers to the action of looking at "what is being equated," in this case, which is the scores. Test score equating is a statistical process that attempts to produce scores that are considered comparable across different test forms, making them interchangeable (Dorans & Cook, 2016). Equating practices often involve not only the choice of statistical equating procedures but also considerations of practical issues related to the use and interpretation of equating results. Scores from two different tests for two or more different groups can be compared if the items are the same and based on the same scale (Kolen & Brennan, 2004).

In large-scale testing programs, the development of equivalent test forms is crucial. Testing programs often produce multiple versions of the same test. At some point, equating several test forms can be done while developing the test itself. However, usually, the difficulty levels of test forms may vary. The requirement of equality must ensure indifference to the test taker, allowing test takers to choose either test form that has been equated without issue (von Davier, 2010). This requirement indicates that once two test forms have been equated, it should not matter to test takers which form they take because the expected scores should be the same for both equated forms (Kolen & Brennan, 2004).

The requirement of equality, though important in theory, is almost impossible to observe in practice because individuals have different opportunities to learn specific test content. Most test takers would prefer a test composed of familiar content over one with unfamiliar content. This definition of equality is based on a clear concern for fairness in equating. If two distributions differ, test takers may be disadvantaged by taking one test rather than the other. For example, test takers with high ability and higher variance in observed scores on test Y compared to test X have a greater risk of not passing a certain score threshold on the first test compared to the second test (von Davier, 2010). Therefore, content balancing in testing is crucial in any attempt to achieve the indifference requirement demanded by this principle. In conclusion, the principle of equality arises due to concerns about equity in equating regarding more than one test form (differences in scores between test takers who took different tests cannot directly imply differences in their abilities, as the difficulty levels of the used test forms would affect these differences).

b. Population Invariance Principle

The population invariance principle refers to a concept in statistics where the observed relationships or patterns between certain variables in an analysis remain constant regardless of changes or variations in the population groups being observed. In other words, this principle states that if a relationship or pattern is found in a statistical analysis in one population group, that relationship or pattern should hold in other population groups, even if those groups have different characteristics.

Equating should be population invariant, meaning that the results of the equating process should not heavily depend on specific subpopulations within the tested population (Dorans & Cook, 2016). In other words, the equating method used should consistently produce results that approximate or can at least be compared to the equating function of the entire population, regardless of variations within subpopulations.

If equating methods do not adhere to the population invariance principle, there can be situations where scores obtained by individuals from certain subpopulations may have distortions or significant differences in their interpretations when compared to individuals from other subpopulations or the entire population (von Davier, 2010). The fundamental difference between classical measurement and modern measurement lies in score invariance, where modern measurement scoring is invariant (does not change) to concerning test items and test takers (Sudaryono, 2011).

The population invariance principle emphasizes that statistical analysis results obtained from one population group should not be seen as unique to that group. This is because variations in population groups can be caused by various factors such as cultural differences, environmental factors, demographics, or other characteristics. By applying the population invariance principle, researchers must be cautious when generalizing results from a statistical analysis of one population group to other population groups. This encourages a more accurate and objective scientific approach, where identified relationships are considered stronger if they consistently hold across different population groups.

c. Symmetry Principle

Equating is a fundamental aspect of educational and psychological measurement. It is done to ensure that scores from different test forms can be meaningfully compared. At the core of this process is the principle of symmetry. Symmetry is the concept that the transformation used to map scores from one test form to another can be reversed. In other

words, scores from Form X to Form Y should be as valid as changing scores from Form Y to Form X (Kolen & Brennan, 2004). Symmetry is a critical principle in equating because non-compliance with it can result in assessment bias and other flaws.

Symmetry, as a fundamental concept in equating, ensures that the equating transformation preserves the ability to compare scores across different test forms. When the symmetry principle is applied, the equating process is not dependent on specific test labels (X or Y), and transformations can be made in both directions without sacrificing score validity. This principle emphasizes the objectivity and fairness of equating procedures, enabling a fair comparison of individual performance or group-level assessment (Kolen & Brennan, 2004; Syahrul et al., 2016).

In the field of assessment equating, the Symmetry Principle, although different from other equating criteria, plays a crucial role often related to standard definitions. For example, the common definitions of linear and equipercentile equating functions inherently ensure adherence to the principle or property of symmetry. Some other groups of scholars consider symmetry to refer to the fact that, regardless of which test is used as the reference or base for transformation, the transformation should be the same. This means that the interpretation of test scores should be the same based on equating from Test A to Test B or vice versa (Kolen & Brennan, 2004; Felan, 2002). However, the intention is that if linear or equipercentile equating methods have been correctly used, the symmetry principle should naturally be satisfied.

Ensuring the symmetry principle is adhered to is crucial in uncovering misconceptions in test equating approaches. Some individuals may sometimes confuse prediction and equating, where a trend in equating in educational assessment is only done through regression methods. This should be noted because using regression as an equating technique is incorrect. However, the requirement of symmetry does not appear to have a fundamental status and is rarely a determining factor in the selection of one linking function over another (Dorans & Cook, 2016).

Statistical methods used to demonstrate the symmetry principle in assessment equating generally involve both Classical Test Theory (CTT) and Item Response Theory (IRT), with the choice depending on the context and availability of specific data. CTT focuses on the properties of observed test scores and can be used to examine the equatability of scores across different test forms through techniques such as equipercentile equating or linear equating. On the other hand, IRT offers a more sophisticated framework that models the relationship between test-taker ability and item difficulty, allowing for more accurate equating through methods like the Stocking-Lord method or Stocking and Lord method (Kolen & Brennan, 2004; Dorans, 2004).

Computer programs used for equating analysis include MS Excel or R Studio, depending on the complexity of the equating task and the researcher's familiarity with these tools. MS Excel provides a user-friendly interface and is suitable for basic equating tasks, especially when dealing with CTT-based equating. In contrast, R Studio is a versatile statistical software suitable for IRT-based equating and complex equating design (Kolen & Brennan, 2004), offering various packages and functions designed specifically for equating analysis.

d. Unidimensionality Principle

The term "unidimensionality" is often used in research contexts to describe the characteristics of test items or test scores (Ziegler & Hagemann, 2015). Unidimensionality refers to the concept that items in a measurement scale should represent the same latent

variable (Hagell, 2014), which cannot be directly measured. The importance of unidimensionality in the items that make up test scores is highly significant in the context of assessments that rely on these scores. Without proper testing to ensure unidimensionality, there is an increased risk of misinterpreting test scores that only reflect one dimension (Ziegler & Hagemann, 2015).

There are at least three reasons why it is important to consider the concept of unidimensionality. First, unidimensionality is the fundamental assumption in calculating valid total scores, according to both classical and modern test theories. Second, clear interpretation requires scores that represent a well-defined single attribute. In this case, scores on a measurement scale designed to measure a single variable should not be overly influenced by variations in one or more other variables. Third, if scores do not reflect a single concept, it becomes difficult to compare two individuals with similar scores (Smith Jr, 2002).

The unidimensionality principle in the context of test equating refers to the principle that equated tests should exclusively measure one specific dimension or characteristic. This implies that the construct or aspect being measured by the test should be unified and not encompass different dimensions. The unidimensionality assumption is met when each test item only measures one test taker's ability. For example, in a mathematics test, the items contained in the test should only measure the test taker's mathematical abilities and not mix in other aspects such as language. However, in practice, it is often challenging to produce test items that truly measure only one ability, as factors like cognitive factors, personality, and environmental factors, such as anxiety, motivation, and guessing tendencies, come into play. Therefore, unidimensionality in tests can be observed only if there is a dominant component that measures the subject's achievement (Sarea & Ruslan, 2019).

Hence, testing the unidimensionality assumption becomes critical. Commonly used methods to test unidimensionality include factor analysis, principal component analysis, dimensionality assessment tests, parallel analysis, Rasch model tests, and internal consistency tests (Smith Jr, 2002). According to Naga (1992), one way to test whether the unidimensionality assumption is met (for the development of a valid and reliable test instrument) is by using factor analysis. Testing unidimensionality can be done with several statistical analyses commonly used in factor analysis, including: (1) Correlation Matrix, which measures the strength of the linear relationships between the variables; (2) Bartlett's Test of Sphericity, which tests whether the correlation matrix is spherical, meaning that variables are independent. In the context of factor analysis, we want significant relationships between variables; (3) Kaiser-Meyer-Olkin (KMO) Test, which measures the suitability and adequacy of data for factor analysis. Higher KMO values indicate that data is suitable for factor analysis; and (4) Eigenvalue Analysis (DIMTEST), to determine how many factors should be extracted. Eigenvalues measure how much variance is explained by each factor (Retnawati, 2016).

Commonly used software programs for testing the unidimensionality assumption in factor analysis or testing assumptions can include: (1) SPSS (Statistical Package for the Social Sciences) (Ardiyaningrum et al., 2018); (2) R; (3) IBM SPSS Amos; (4) SAS (Statistical Analysis System); (5) JASP; (6) Mplus; and (7) Python. The results of these tests are crucial as they affect the validity and reliability of test results. Although in practice, the unidimensionality assumption is very difficult to fully satisfy due to cognitive, personality,

administrative, and test implementation factors (Sudaryono, 2011), testing this assumption remains a key step in analyzing tests using IRT.

Therefore, testing the unidimensionality assumption is a primary concern in the development of measurement instruments in various educational fields (Margono, 2013), including determining the appropriate model for test result analysis (Hoe, 2008). Additionally, testing the unidimensionality assumption is also relevant in detecting Differential Item Functioning (DIF) through an item response theory approach (Samritin, 2022). In the context of education and ability measurement, the use of Item Response Theory (IRT) assuming unidimensionality has become a vital tool in evaluating measurement instruments (Meijer & Tendeiro, 2018), with high validity and reliability (Friyatmi, 2018; Muhdar, 2023; Syamsuddin, 2023). Therefore, ensuring unidimensionality is a crucial step in producing effective and accurate measurement instruments.

D. CONCLUSION AND SUGGESTIONS

Assessment is a crucial aspect in the field of education and performance evaluation across various sectors. Fundamental principles in assessment, such as the principles of equity, population invariance, symmetry, and unidimensionality, need to be effectively applied to ensure fair and reliable assessments. The principle of equating has become increasingly important in the context of globalization and high mobility, where comparisons and recognition of assessment results have significant implications. Test score equating is a statistical process used to understand the relationship between scores on two or more tests, enabling fair comparisons of individual performance or group-level assessments. However, despite the importance of equating principles in assessment, challenges remain in developing accurate equating methods, utilizing technology to support the assessment process, and implementing policies that promote equity in assessment.

To enhance the quality of assessment, several recommendations can be considered,: First, educators and researchers should carefully understand and apply the fundamental principles of assessment, especially equating principles, to ensure the fairness and reliability of assessments. Second, the use of technology in assessment should be enhanced to improve the efficiency and accuracy of the assessment process. Third, the importance of testing the assumption of unidimensionality in tests should be further emphasized to ensure that tests genuinely measure a specific dimension or characteristic. Fourth, researchers should continue to develop more accurate and context-relevant equating methods in the current assessment context.

ACKNOWLEDGEMENT

We would like to express our gratitude to the Teaching Team of the Measurement and Assessment Course in the Doctoral Program of Research and Evaluation in Education at Universitas Negeri Yogyakarta.

REFERENCES

Aminah, N. S. (2012). Karakteristik metode penyetaraan skor tes untuk data dikotomos. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *16*, 88–101. http://dx.doi.org/10.21831/pep.v16i0.1107

Ardiyaningrum, M., Kusuma, C., & Trisniawati, T. (2018). Analisis Butir Try Out Ujian Nasional Matematika Sekolah Dasar Di Daerah Istimewa Yogyakarta Tahun 2017. *Taman Cendekia: Jurnal Pendidikan Ke-SD-An*, *2*(2), 206–211. https://doi.org/10.30738/tc.v2i2.2819

Bayo-Moriones, A., Galdon-Sanchez, J. E., & Martinez-de-Morentin, S. (2020). Performance appraisal: dimensions and determinants. *The International Journal of Human Resource Management*, *31*(15), 1984–2015. https://doi.org/10.1080/09585192.2018.1500387

Blanco-Vogt, A., & Schanze, J. (2014). Assessment of the physical flood susceptibility of buildings on a large scale–conceptual and methodological frameworks. *Natural Hazards and Earth System Sciences*, *14*(8), 2105–2117. https://doi.org/10.5194/nhess-14-2105-2014

Cassidy, S. (2007). Assessing 'inexperienced'' ability to self-assess: Exploring links with learning style and academic personal control.' *Assessment & Evaluation in Higher Education*, *32*(3), 313–330. https://doi.org/10.1080/02602930600896704

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227–246. https://doi.org/10.1177/0146621604265031

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. Taylor & Francis.

Duarte, M. E., & Rossier, J. (2008). Testing and assessment in an international context: Cross-and multi-cultural issues. In *International handbook of career guidance* (pp. 489–510). Springer. https://doi.org/10.1007/978-1-4020-6230-8_24

Elkhoury, E., Ali, A., & Sutherland-Harris, R. (2023). Exploring Faculty Mindsets in Equity-Oriented Assessment. *Journal of University Teaching & Learning Practice*, *20*(5), 13. https://doi.org/10.53761/1.20.5.12

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists multivariate*. London, UK: Erlbaum Publishers.

Erfianti, L., Istiyono, E., & Kuswanto, H. (2019). Developing lup instrument test to measure higher order thinking skills (HOTS) Bloomian for senior high school students. *International Journal of Educational Research Review*, *4*(3), 320–329. https://doi.org/10.24331/ijere.573863

Felan, G. D. (2002). *Test Equating: Mean, Linear, Equipercentile, and Item Response Theory.*

Friyatmi, F. (2018). Estimasi parameter tes dengan penskoran politomus menggunakan graded response model pada sampel kecil. *Jurnal Inovasi Pendidikan Ekonomi (JIPE)*, *8*(1), 22–31. https://doi.org/10.24036/01104490

Gaytan, J., & McEwen, B. C. (2007). Effective online instructional and assessment strategies. *The American Journal of Distance Education*, *21*(3), 117–132. https://doi.org/10.1080/08923640701341653

Gronlund, N. E. (2003). Assessment of Student Achievement. 7. útgáfa. *Bandaríkin: Allyn and Bacon*.

Hagell, P. (2014). Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: the primacy of theory over statistics. *Open Journal of Statistics*, *4*(6), 456–465. https://doi.org/10.4236/OJS.2014.46044

Heinrichs-Graham, E., Walker, E. A., Taylor, B. K., Menting, S. C., Eastman, J. A., Frenzel, M. R., & McCreery, R. W. (2022). Auditory experience modulates fronto-parietal theta activity serving fluid intelligence. *Brain Communications*, *4*(2), fcac093. https://doi.org/10.1093/braincomms/fcac093

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899%2FJCE-18-22

Hoe, S. L. (2008). Issues and procedures in adopting structural equation modelling technique. *Journal of Quantitative Methods*, *3*(1), 76. https://ink.library.smu.edu.sg/sis_research/5168

Humphry, S. (2006). The impact of differential discrimination on vertical equating. *ARC Report*.

Hung, Y. (2019). Bridging assessment and achievement: Repeated practice of self-assessment in college English classes in Taiwan. *Assessment & Evaluation in Higher Education*, *44*(8), 1191–1208. https://doi.org/10.1080/02602938.2019.1584783

Immonen, K., Oikarainen, A., Tomietto, M., Kääriäinen, M., Tuomikoski, A.-M., Kaučič, B. M., Filej, B., Riklikiene, O., Vizcaya-Moreno, M. F., & Perez-Canaveras, R. M. (2019). Assessment of nursing students' competence in clinical practice: a systematic review of reviews. *International Journal of Nursing Studies*, *100*, 103414.

https://doi.org/10.1016/j.ijnurstu.2019.103414
Johnson, S. N., Gallagher, E. D., & Vagnozzi, A. M. (2021). Validity concerns with the revised study process questionnaire (R-SPQ-2F) in undergraduate anatomy & physiology students. *Plos One*, *16*(4), e0250600. https://doi.org/10.1371/journal.pone.0250600
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*.
Lucey, C. R., Hauer, K. E., Boatright, D., & Fernandez, A. (2020). Medical education's wicked problem: achieving equity in assessment for medical learners. *Academic Medicine*, *95*(12S), S98–S108. https://doi.org/10.1097/acm.0000000000003717
Lyngsø, A. M., Godtfredsen, N. S., Høst, D., & Frølich, A. (2014). Instruments to assess integrated care: a systematic review. *International Journal of Integrated Care*, *14*. https://doi.org/10.5334/ijic.1184
Mardapi, D. (2017). Pengukuran Penilaian dan Evaluasi Pendidikan Edisi 2. *Yogyakarta: Parama Publishing*.
Margono, G. (2013). Aplikasi analisis faktor konfirmatori untuk menentukan reliabilitas multidimensi. *Statistika*, *13*(1). https://doi.org/10.29313/jstat.v13i1.1069
Mau, S. (2020). Numbers matter! The society of indicators, scores and ratings. *International Studies in Sociology of Education*, *29*(1–2), 19–37. https://doi.org/10.1080/09620214.2019.1668287
McDonald, M. (2002). *Systematic assessment of learning outcomes: Developing multiple-choice exams*. Jones & Bartlett Learning.
Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 413–443. https://doi.org/10.1002/9781118489772.ch15
Muhdar, R. (2023). Assesmen Kompetensi Minimum Numerasi Program Merdeka Belajar. *Jurnal Ilmiah Wahana Pendidikan*, *9*(12), 407–411. https://doi.org/10.5281/zenodo.8079162
Naga, D. S. (1992). Pengantar teori sekor pada pengukuran pendidikan. *Jakarta: Gunadarma*.
Nayir, F., Brown, M., Burns, D., Joe, O., Mcnamara, G., Nortvedt, G., Skedsmo, G., Gloppen, S. K., & Wiese, E. F. (2019). Assessment with and for migration background students-cases from Europe. *Eurasian Journal of Educational Research*, *19*(79), 39–68. https://doi.org/10.14689/ejer.2019.79.3
Papastephanou, M. (2005). Globalisation, globalism and cosmopolitanism as an educational ideal. *Educational Philosophy and Theory*, *37*(4), 533–551. https://doi.org/10.1111/j.1469-5812.2005.00139.x
Popham, W. J. (2008). *Transformative Assessment: Association for Supervision and Curriculum Development. 1703 North Beauregard Street, Alexandria, VA 22311-1714*. Tel.
Putri, F. S., & Istiyono, E. (2017). The Development of Performance Assessment of STEM-Based Critical Thinking Skill in the High School Physics Lessons. *International Journal of Environmental And Science Education*, *12*(5), 1269–1281. http://www.ijese.net/makale/1894.html
Retnawati, H. (2016). Validitas reliabilitas dan karakteristik butir (Validity, reliability and item charactheristic). *Yogyakarta, Indonesia: Parama Publishing*.
Retnawati, Heri. (2014). Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana. *Yogyakarta: Nuha Medika*.
Samritin, S. (2022). Identifikasi Muatan Differential Item Functioning Pada Data Ujian Nasional Matematika. *Journal on Education*, *4*(4), 1675–1684. https://doi.org/10.31004/joe.v4i4.2508
Sarea, M. S., & Ruslan, R. (2019). Karakteristik Butir Soal: Classical Test Theory Vs Item Response Theory? *Didaktika: Jurnal Kependidikan*, *13*(1), 1–16. http://dx.doi.org/10.30863/didaktika.v13i1.296
Scott, E. E., Wenderoth, M. P., & Doherty, J. H. (2019). Learning progressions: An empirically grounded, learner-centered framework to guide biology instruction. *CBE—Life Sciences Education*, *18*(4), es5. https://doi.org/10.1187/cbe.19-03-0059

Shofwanthoni, M. A., Ridlo, S., & Elmubarok, Z. (2019). The development of authentic assessment instrument of Hajj Manasik practices of IX grade of SMP PGRI 10 Candi in Sidoarjo Regency. *Journal of Research and Educational Research Evaluation*, *8*(1), 14–21. https://doi.org/10.15294/jere.v8i1.28361

Singh, J., Steele, K., & Singh, L. (2021). Combining the best of online and face-to-face learning: Hybrid and blended learning approach for COVID-19, post vaccine, & post-pandemic world. *Journal of Educational Technology Systems*, *50*(2), 140–171. https://doi.org/10.1177/00472395211047865

Smith Jr, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231. https://europepmc.org/article/med/12011501

Sudaryono, S. (2011). Implementasi Teori Responsi Butir (Item Response Theory) Pada Penilaian Hasil Belajar Akhir di Sekolah. *Jurnal Pendidikan Dan Kebudayaan*, *17*(6), 719–732. https://doi.org/10.24832/jpnk.v17i6.62

Syahrul, Mansyur, & Rosdiyanah. (2016). Pengaruh Jumlah Butir Anchor Terhadap Hasil Penyetaraan Tes Berdasarkan Teori Respon Butir. *Jurnal Kependidikan*, *46*(2). http://dx.doi.org/10.21831/jk.v46i2.10935

Syamsuddin, S. (2023). Implementasi Classic Test dan Item Respon Theory Pada Penilaian Tes Pembelajaran Matematika. *EDUSCOPE: Jurnal Pendidikan, Pembelajaran, Dan Teknologi*, *8*(2), 28–43. https://doi.org/10.32764/eduscope.v8i2.3488

Umar, A.-T., & Majeed, A. (2018). The Impact of Assessment for Learning on Students' Achievement in English for Specific Purposes: A Case Study of Pre-Medical Students at Khartoum University: Sudan. *English Language Teaching*, *11*(2), 15–25. http://doi.org/10.5539/elt.v11n2p15

von Davier, A. (2010). *Statistical models for test equating, scaling, and linking*. Springer Science & Business Media.

Wahhab, K. A., & Rizko, N. J. (2019). The importance of evaluating the environmental design and performance of student projects as a product of architecture departments: A case study. *Periodicals of Engineering and Natural Sciences*, *7*(3), 1286–1299. http://dx.doi.org/10.21533/pen.v7i3.666

Were, M. C., Sinha, C., & Catalani, C. (2019). A systematic approach to equity assessment for digital health interventions: case example of mobile personal health records. *Journal of the American Medical Informatics Association*, *26*(8–9), 884–890. https://doi.org/10.1093/jamia/ocz071

Wiliam, D., & Thompson, M. (2017). Integrating assessment with learning: What will it take to make it work? In *The future of assessment* (pp. 53–82). Routledge.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. In *European Journal of Psychological Assessment*. Hogrefe Publishing.