JURNAL PENDIDIKAN: RISET&KONSEPTUAL

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

Meta-Analysis of Item Generation Procedures Used in Selected Standardised Tests in Education

Oyeronke Christiana Paramole[(1)], Eyitayo Rufus Ifedayo Afolabi[(2)]

[1]Department of Social Studies Education
Al-Hikmah University, Ilorin, Nigeria
[2]Department of Educational Foundations and Counselling
Obafemi Awolowo University, Ile-Ife, Nigeria
*Corresponding author: olaronkus27@gmail.com

Abstract: Meta-analysis is a systematic process combining multiple studies' results to create a comprehensive understanding of a research question. It uses statistical methods to synthesise the findings of different studies, offering a quantitative summary of the overall effect. The study analysed the differences in studies on validating standardised tests in education using various procedures. The meta-analysis of 130 empirical studies on the validation of standardised educational tests published between 1988 and 2017 revealed that authors predominantly rely on literature reviews for item generation, followed by theory reviews and expert reviews. The data was analysed using percentages and counts. The majority of studies (80%) used literature reviews to generate ideas, with a smaller percentage (60%) and (38%) for theory and expert reviews. The majority of authors used Guttman divided half and internal consistency, with Alpha if item removed being the most commonly used method (86.2%). Construct validity was the primary method used (89.2%). The factorability of the correlation matrix was the main reason for the Barlett's test of sphericity score (75.4%). Most authors reported factor retention (60.8%) and used the Unweighted Least Squares (51.5%) approach for factor extraction. The most commonly reported rotation technique was orthogonal (Varimax) (59.2%). In conclusion, the text provides a comprehensive overview of test validation, emphasising the importance of establishing validity for ensuring trustworthy and meaningful test results. Future research should continue to explore the complexities of test validation, seeking to improve practices and develop more standardised approaches that ensure the reliability and validity of tests across various contexts.

Tersedia Online di
http://journal.unublitar.ac.id/pendidikan/index.php/Riset_Konseptual

Sejarah Artikel
Diterima pada : 15-10-2024
Disetujui pada : 05-10-2024
Dipublikasikan pada :22-10-2024

Kata Kunci:
Education, Item Generation, Meta-Analysis, Procedures, Standardised Tests.

DOI:
http://doi.org/10.28926/riset_konseptual.v8i4.1111

INTRODUCTION

Standardised tests are crucial in education for making important decisions about curriculum, placement, and instruction. Validating these tests is essential to ensure their validity and reliability. However, the methods used in validating these tests can vary significantly across research studies, leading to inconsistencies and potential inaccuracies. This study conducts a comprehensive literature review of methods used for validating standardised tests from 1988-2017, identifying the strengths and weaknesses of each method. This review provides a valuable overview of current practices and identifies areas for future research and improvement in standardised test validation. Jiang et al. (2017) emphasize the importance of a uniform fashion in administering and scoring standardized tests for all test takers. These tests should reflect the talents or skills being tested, rather than high-stakes, time-constrained examinations. Standardised examinations can cover various subjects such as aptitude, driving tests, creativity, personality, professional ethics, and IQ. This ensures a consistent and accurate evaluation of test takers.

Cappella et al. (2016) highlight the fundamental elements of standardized tests, which consist of a set of inquiries and a summative score. These tests can be aptitude tests, personality tests, intelligence tests, and more. Zumbo et al. (2002) highlight aptitude tests as assessing a person's learning capacity while Gliner et al (2011)

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

emphasize the importance of interval level measurement in educational research. These tests provide more precise data on human behavior, aiding decision-making and allowing for more accurate deductions and conclusions. Personality tests, such as emotion, attitude, and motivation, examine how the test-taker prefers to behave in specific situations. Overall, standardized tests are essential for evaluating a person's potential and capabilities. Research on the validation of standard tests has been expanding, and this expansion calls for accurate and reliable assessments of phenomena that are pertinent to its methods. It is difficult to validate tests for subjective concepts like decision-making and quality of life that are important in ethical problem-solving. However, inadequate construct definition reduces psychometric qualities, and improper item reduction can have a range of unanticipated effects, such as issues with validity and skewed results. Standardised exams are vulnerable to validity due to the effects of central tendency, leniency, harshness, hallow effect, and the difficulty of faking (Smith, 2013).

Ogbonnaya (2019) explored the reliability of tests based on measurement mistakes. They questioned if the same test findings would be the same if given to the same respondents under the same circumstances. The more measurement errors, the less reliable the test is. The more errors, the more trustworthy comparisons can be established. The unreliability of an assessment increases with the number of errors identified. Validity was defined by Whiston (2009) as the extent to which theory and evidence support how the intended test user would perceive test results. Similarly, Kaplan and Saccuzzo (2001) saw validity as the proof used to conclude a test result. The degree to which a test measures what it is intended to assess is known as validity. When a test is validated, analytical methods are used to make sure that only things that meet specific requirements are included in the test's final edition. A test must be trustworthy to be valid, yet a test might be reliable without being valid. For instance, a test that is calibrated wrongly can produce identically false weight readings.

According to Kane (2013), validating pertinent standardised tests is a challenging and time-consuming task. However, a variety of processes relating to the development and validation of items have changed throughout time to provide legitimate standardised examinations. While some, including self-report tests, have universal principles, underlying assumptions, and have an impact on the calibre of tests generated, some of these methods are context-specific. Haladyna and Rodriguez (2013) and Na-Nan et al. (2018) emphasize the importance of item development, scale development, and scale evaluation in validating tests. They use seven stages to characterize the testing process: establishing the construct, designing the test, creating an item pool, giving the test, reviewing results, computing coefficient alpha, and factor analysis. Barry et al. (2011) outline eight steps before factor analysis, including defining the outcome, creating the item pool, choosing the measure format, having a professional review, considering validation items, administering tests to a development sample, evaluating the test, and optimizing scale length. Rossiter (2010) uses construct definition, object categorisation, attribute classification, rater identification, scale construction, enumeration, and reporting for test validation methods.

Olaleye (2015) and Adeyemo and Olatomide (2017) are two researchers who have developed standard tests to measure judgements, views, perceptions, and sentiments. Olaleye (2015) created a teacher disposition scale using five steps: defining components, developing items, pilot testing, implementing the scale, and analyzing results. Adeyemo and Olatomide (2017) created a retirement adjustment scale for retired secondary school teachers in southwest Nigeria, using four test development procedures: item development, scale development, pilot study, and scale analysis. These methods raise concerns about the validity and reliability of standardized tests, highlighting the need for more comprehensive and effective methods. Test creators and researchers employ varying methods based on their customs, expertise, and resources. Glass (2012) divided meta-analysis into three categories: primary, secondary and meta-analysis. Primary analysis involves the

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

analysis of an original study by the researcher, secondary analysis uses statistics to understand the issue covered, and meta-analysis analyzes numerous empirical studies conducted in the past (Sherif, 2018). Meta-analysis is a systematic summarising of studies to gather information from many studies, yielding statistical indications for broad inferences (Hartung et al., 2011). It entails aggregating these estimates across studies to provide a summary by characterizing the outcomes of each trial with a numerical measure of impact magnitude, such as a correlation coefficient, a standardized mean difference or an odds ratio.

Burke and Landis (2013) and Oswald and McCloy (2013) have both contributed to the field of meta-analysis, a statistical method that combines the findings of multiple studies to provide a composite perspective on the mean influence of a construct. This approach, which focuses on merging results from other research that addresses the same concerns, is crucial in ensuring the reliability of findings. Wiernik and Dahlke (2020) argue that meta-analysis applies to a wide range of investigations and uses unbiased formulae, such as Fisher, Winder, and Steffer combination tests, to summarize the outcomes of each study using a numerical index of effect size. This approach helps in drawing firm conclusions when there are multiple studies involved, ensuring the validity of the results. According to Iacobucci et al. (2023), effect size refers to the degree to which a null hypothesis is false and is used to examine the relationship between independent and dependent variables in a study. In meta-analysis, the intensity and direction of the link between variables are known as effect size. The standard deviation of either group is used to split the difference between the means of each pair of treatment conditions. The goal is to more accurately estimate the genuine effect size compared to an effect size generated from single research under specific assumptions and conditions. The term is widely used in education and is occasionally used to describe relationships or the standardised mean difference.

Hedges et al. (2010) identified three common computations for determining effect sizes: correlation coefficients, averages, and proportions. These methods calculate effect magnitude in various ways, such as using averages for numerical results, proportions for nominal results and correlations for link-based results. Effect sizes and their confidence intervals are often displayed in a funnel plot (Sedgwick & Marston, 2015). Heterogeneity tests are used to determine if the effect size distribution conforms to the normal distribution (Huedo-Medina et al., 2006). The Q-statistics, with chi-square distribution (k-1) degrees of freedom, is used to determine if an achieved value of Q is significant enough to demonstrate heterogeneity at a specific level. Publication bias is a significant issue in meta-analyses, where findings must be published in the same format as other scientific papers (Hussain et al., 2019). Research with positive results and a large sample size is more likely to be published and appear in searches. Trinquart et al. (2018) suggest that a forest plot can be used to show if the meta-analysis missed studies, it should have been included. However, there are discrepancies and disagreements among methods used in the validation of standardised tests, leading to the need for a comprehensive summary of prior empirical studies on these methods.

To guarantee that standardised tests are accurate and valid, both qualitative and quantitative processes are used in their preparation. The quantitative methods are mostly at the stage of item analysis and scale validation, whereas the qualitative procedures are primarily in the stage of creating initial test items. The methods used to validate standardised tests vary from researcher to researcher. The a priori techniques may vary or have flaws, which might result in construct metrics with questionable validity. Tests that produce measurements that are insufficiently reflective of the constructs they are supposed to measure may lead to inaccurate or misleading results and unsuitable judgments. These choices might hurt a person's health, academic performance, profession, personal issues, personality, social interactions, and skill development, among other things. Additionally, from the perspective of research, researchers would distribute incorrect information resulting from incorrect methods, and

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

those who reproduce similar testing validation procedures with incorrect lines are likely to confirm the alleged flaws. This study identifies the most effective methods for validating standardised exams in education (1988-2017) by examining empirical findings and combining statistical indicators. The research question is: What are the different procedures used to verify standardized tests in the selected studies? The study's unique goal is to determine the most effective approaches for validating standardized tests.

The increasing awareness of disparities in educational outcomes among different demographic groups necessitates a meta-analysis of item generation procedures to identify if existing tests inadvertently favor certain groups over others. This research is urgent as educational stakeholders strive to create assessments that are fair and equitable, thus promoting a more just educational landscape. As curricula and educational standards shift, so too must the methods used to evaluate student learning. Item generation procedures must be flexible enough to adapt to contemporary educational demands, including the integration of technology and the emphasis on critical thinking and problem-solving skills. Research in this area can provide valuable guidance on how to effectively update item generation methodologies to align with current trends in education. Enhancing test security and integrity is essential for maintaining the credibility of assessments. Research into best practices for generating secure and robust test items can inform the development of more resilient assessment systems, particularly in high-stakes testing environments. Informing policy and practice is crucial, as educational policymakers rely on the findings from standardized tests to make informed decisions regarding curriculum development, resource allocation, and educational interventions. Conducting a meta-analysis of existing methodologies can provide policymakers with evidence-based recommendations, ensuring that assessments are effective and aligned with educational goals.

This research is a comprehensive examination of item generation procedures across diverse standardised tests, providing a more holistic understanding of current practices. It analyzes multiple tests, ranging from elementary assessments to advanced placement exams, to identify trends, best practices, and gaps in methodology. The study also conducts a comparative analysis of various item generation procedures, categorizing them into distinct frameworks such as expert review, empirical data analysis, and automated item generation. This approach helps in identifying the strengths and weaknesses of each method and provides insights into how they can be optimally integrated to enhance the quality of assessments. The research emphasizes the importance of aligning assessment tools with pedagogical objectives to ensure valid and reliable indicators of student performance. The study also incorporates recent technological advancements in item generation, such as artificial intelligence and machine learning, into standardised testing, contrasting traditional methods with innovative approaches that leverage data analytics and algorithmic processes. The research places significant emphasis on equity and accessibility within the context of item generation, investigating how different procedures account for diverse student populations, inclusive of varying linguistic, cultural, and educational backgrounds. The findings reveal that while some item generation processes incorporate inclusive practices, others may inadvertently perpetuate biases or disadvantage certain groups of students. The research also outlines several recommendations for future studies, including further exploration into the longitudinal impacts of different item generation methods on student learning outcomes and advocating for collaborative efforts among educators, psychometricians, and technology developers to create more effective and inclusive assessment tools.

METHOD

An ex-post-facto research design was used in the study. It is a technique for gathering potential causes of already-happened events that cannot be changed. The methodology was chosen to compile research on techniques for the validation of

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

educational standardised examinations, over which the researcher has no influence. To find research, the ancestor strategy to information retrieval was also applied. By following citations from one research to another through bibliographies mentioned in earlier studies, the ancestry technique, assures that similar studies are discovered (Ovute, 2015).

Population

The population of the study consisted of all prior empirical studies on methods for validating standardised tests that had been published in national and international journals between 1988 and 2017 with the main objective of validating standardised tests.

Sample and Sampling Technique

Due to the limited number of research that had been done in this field locally, the sample size was 130 empirical investigations that had previously been undertaken on the validation of standardised tests both locally and globally. The research investigations included unpublished theses and dissertations, as well as published and unpublished journal papers. The sample was selected using the following standards:

- The study made use of quantitative methods in reporting the analysis of the results of the tests developed.
- The test statistics are convertible to Pearson-r.
- The study was carried out or published between 1998 and 2017.
- The study reported a significant level and sample size of its result.

The distribution of the studies to be used for the meta-analysis in terms of publication source is as indicated in Table 1

Table 1 The Distribution of 130 Studies that were Used for the Research

Type of Study	Number of Studies
Journal (Published) International	73
Journal (Published) Local	35
Ph.D. Theses (Unpublished)	5
Masters Theses (Unpublished)	17
Total	130

Source: Author's search 2024

Research Instrument

A coding sheet modified by Adeyemo (2012), Subar (2014), and Odeyemi (2016) was adopted and used for the selection of eligible studies.

Techniques for Data Analysis

The major computation in meta-analysis is effect size, which gauges how strongly and in what direction a link between two variables is (Cafri et al., 2010). The data was analysed using percentages and counts using the Statistical Product and Service Solutions (IBM SPSS) from International Business Machines.

Results and Discussion

To answer the research question, procedures for the validation of test items were highlighted and percentages and counts of data coded for each study were calculated using the SPSS package. The procedures include the generation of items, methods of item deletion, forms of reliability, types of validity, factor analytic procedure, factor ability of correlation matrix, analysis of factor retention, factor extraction method and rotation method.

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

Table 2 The Frequency Counts and Percentage of 130 Studies Selected on Validation of Standardised Tests

Type of Study	Number of Studies	Percentage
Articles (Published) International	73	56.12
Articles (Published) Local	35	26.92
PhD Theses (Unpublished)	5	3.84
Masters Theses (Unpublished)	17	13.08
Total	130	100

Source: Author's analysis 2024

Table 3 Frequencies and Percentages of Procedures for Standardised Tests Validation

S/N	Method of test validation	Characteristics	Frequency N=130	Percentage %
1.	Generation of item	Theory	52	60.0
		Interview/Focus group	18	13.8
		Experts review initial item	44	38.8
		Literature review	105	80.8
		Clinical / Observation	25	19.2
		Responses to open ended questions	26	20.0
2.	Methods of item deletion	Theory	10	7.7
		Advice	8	6.2
		Item total correlation	29	22.7
		Alpha if item deleted	112	86.2
		Factor Loading	106	81.5
		Item, means, standard deviation	50	60.8
		Item correlation too low	43	33.1
		Content coverage	29	22.7
		Loading on wrong factor	27	22.0
		Badly worded	42	32.3
		Low variance	36	27.7
3.	Forms of reliability	Internal consistency (Cronbach's Alpha)	123	94.5
		Internal consistency (Spearman Brown split half)	85	64.5
		Internal consistency (Guttman's split half)	123	94.6
		Test retest (test of stability)		
		Inter–rater (Marginal or average)	119	91.5
		IRT		

4.	Types of validity	Construct	116	89.2
		Content	94	72.3
		Predictive	29	22.7
		Discriminant	14	10.8
		Convergent	98	77.3
		Criterion	8	6.2
5.	Factor Analytic Procedure	Exploratory factor analysis	51	39.2
		Confirmatory factor analysis	42	32.3
		EFA and CFA	43	33.1
6.	Factor ability of correlation matrix	Absolute sample size	54	41.5
		Inter–item correlation	39	30.0
		Participant per item ratio	28	21.5
		Bartlett's test of sphericity	98	75.4
		KMO method of sampling adequacy	86	66.2
7.	Analysis of factor retention	Loadings	44	38.8
		Communalities	79	60.8
		Item analysis	42	32.3
		Eigenvalues	77	59.2
		Scree plot	74	56.9
		The minimum portion of variance accounted for by factor	46	35.6
		Parallel analysis	20	15.4
8.	Factor extraction method	PCA Principal component analysis	53	40.8
		Common factor analysis	26	20.0
		Principal axis factoring	11	8.5
		Maximum likelihood	22	16.9
		Unweigheted least squares	67	51.5
9.	Rotation method used	Orthogonal (Varimax)	77	59.2
		Promax	21	16.2
		Oblique	16	12.3
		Oblium	52	40.0

Source: Author's analysis 2024

Table 3 shows the steps followed in the procedures for validating standardised tests in education. Specifically, to achieve this, the criteria for validating test items were examined. Six criteria have been identified for item generation (Worthington & Whittaker, 2006). For each of the criteria examined in the 130 studies, the result showed that the option with the highest frequency count for item generation was literature review at 105 (80.8%) followed by expert review of initial items at 52 (60%). The least method used in the generation of items was interview/focus group discussion (FGD) 18 (13.8%). Items were mostly deleted to increase coefficient alpha in the development of their measure. Method of item deletion: alpha if item deleted at 112 (86.2%) and factor loading at 106 (81.5%) were the most used methods while advice was at 10 (7.7%) and theory at 8 (6.2%) were sparsely used. For the method of reliability, almost all the results of this study showed that authors reported coefficient alpha reliabilities, Cronbach's Alpha and Guttman's split half at 123 (94.5%),

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

Spearman-Brown split half at 85 (64.5%) and inter-rater at 29 (22.7%). Eight studies at 6.2% reported using IRT and only 4 studies reported low reliability of less than 0.55.

Types of validity used: In the vast majority of articles construct validation was reported at 116 (89.2%), also content and convergent validity were reported at 94 (72.3%) and 98 (77.3) respectively. Criterion, descriptive and discriminant validity were sparsely used by authors. Factor analytic procedures used were exploratory factor analysis at 51 (39.2%), confirmatory factor analysis at 42 (32.3), and studies that made use of the two procedures concurrently at 43 (33.1%). Factor ability of correlation matrix, results showed that 54 (41.5%) of the studies used absolute sample size; Inter – item correlation 39 (30.0%); Participant per item ratio 28 (21.5%); Bartlett's test of sphericity at 98 (75.4%) and KMO method of sampling adequacy 86 (66.2%).

Bartlet's test of sphericity was mostly used criteria for the factorability of correlation matrix with 98 (75.42%) instruments of the selected scales reporting its use followed by Kaiser Meyer Olkins (KMO) method of sampling adequacy reporting 86 (66.2%). Participant per item ratio had the lowest percentage of 21.5 usages by 28 authors. Thirty-nine authors made use of Inter-item correlation with 30.0%. The analysis of factor retention was carried out, loadings were 44 (38.8%), commonalities 79 (60.8%), item analysis 42 (32.3%), eigenvalues 77 (59.2%), scree plot 74 (6.9%), minimum portion of variance accounted for by factor 46 (35.6%), parallel analysis 20 (15.4%). Factor extraction method, PCA Principal component analysis was used by 53 authors at (40.8%), common factor analysis was used at 26 (20.0%), principal axis factoring 11 (8.5%), maximum likelihood 22 (16.9%), while unweighted least squares 67 (51.5%). Some authors did not specify the method used in deleting or retaining their items. The rotation method used by each author was also categorised as orthogonal varimax, oblimin, promax and oblique. Orthogonal (varimax) at 77 (59.2%) was the most used method of rotation followed by oblium at 52 (40.0%), promax and oblique had the lowest use at 21 (16.2%) and 16 (12.3%).

DISCUSSION

The criteria for validating test items were examined. This study relies primarily on Clifton (2020) as the most reliable resource. Various procedures used in test validation considered in this study are: generation of item, methods of item deletion, forms of reliability, types of validity, factor analytic procedure, factor ability of correlation matrix, analysis of factor retention, factor extraction method and rotation method used. However, six criteria have been identified for item generation from journals (Worthington & Whittaker, 2006) and also from recommendations in texts and unpublished thesis. For each of the criteria examined in the 130 studies, results showed that the option with the highest frequency count for item generation was a literature review followed by an expert review of initial items. The last method used in the generation of items was an interview/focus group. Akpunne and Akinnawo (2018) did not report how items were generated while few reported more than one criterion. Pett et al. (2003) recommended that researchers should choose the quality of the item pool correctly and plan the approach to the generation of items critically. Items were mostly deleted to increase coefficient alpha in the development of their measure. Methods of item deletion identified in this study were: theory, advice, item-total correlation, alpha if item deleted, factor loading, item means, standard deviation, item correlation too low, content coverage, loading on wrong factor, badly worded and low variance. From the options listed for item elimination, alpha if item deleted and factor loading were the most used methods in the studies selected while advice and theory were sparsely used.

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

CONCLUSION

The meta-analysis of item generation procedures in selected standardized tests in education revealed a diverse range of approaches, from traditional methods like multiple-choice questions to more innovative formats like performance-based assessments and technology-enhanced items. Each method has its advantages and limitations, such as efficient scoring and broad content coverage but not fully capturing higher-order thinking skills. Performance-based assessments offer a more holistic view of student capabilities but often require more complex scoring rubrics and can be resource-intensive. The analysis also highlighted the increasing incorporation of technological tools in item generation, such as automated item generation systems and computer-adaptive testing, which have revolutionized the way assessments are constructed. However, this reliance on technology introduces concerns regarding accessibility, equity, and potential biases embedded in algorithmic processes. Educational policymakers and practitioners must consider broader contextual factors that influence item generation procedures, such as cultural relevance, language diversity, and socioeconomic disparities, to ensure that standardized tests serve all students equitably. The involvement of diverse stakeholder groups, including educators, students, and community representatives, in the item-generation process can lead to more inclusive and representative assessments. Ongoing professional development for educators involved in test design is crucial to foster a deeper understanding of item generation methodologies and their implications for student learning. Collaboration among test developers, researchers, and practitioners is essential to share best practices and innovate in the field of assessment. Future research should focus on the long-term impacts of these methods on educational outcomes, develop guidelines for best practices in item generation that incorporate equity and accessibility considerations, and assess the effectiveness of emerging assessment formats and their implications for diverse learner populations.

REFERENCE

Adeyemo, E. O. & Olatomide, O. O. (2017). Validation of Retirement Adjustment Scale for Retired Teachers of Secondary Schools in Osun State, Nigeria. *International Journal of Education and Research*, *5*, 209-220.

Adeyemo, E. O. (2012). A Meta-analysis of Empirical Studies on the Validity of University Matriculation Examinations in Nigeria. *International Journal of Learning*, *18*(3).

Akpunne, B.C & Akinnawo, O.E. (2018). Validation of smartphone addiction scale–short version on Nigerian university undergraduates. International Journal of Computer Science and Mobile Computing 7(11), 136-141

Barry, A. E., Chaney, E. H., Stellefson, M. L., & Chaney, J. D. (2011). So You Want To Develop A Survey: Practical Recommendations For Scale Development. *American Journal of Health Studies*, *26*(2).

Burke, M. J., & Landis, R. S. (2013). Methodological and conceptual challenges in conducting and interpreting meta-analyses. In *Validity Generalization* (pp. 287-309). Psychology Press.

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, *45*(2), 239-270. https://doi.org/10.1080/00273171003680187

Cappella, E., Aber, J. L., Kim, H. Y., Gitomer, D. H., & Bell, C. A. (2016). Teaching beyond achievement tests: Perspectives from developmental and education science. *Handbook of research on teaching*, *5*, 249-347. https://doi.org/10.3102/978-0-935302-48-6_4

Clifton, J. D. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259. https://doi.org/10.1037/met0000236

Glass, G. V. (2012). Meta-analysis: The quantitative synthesis of research findings. In *Handbook of complementary methods in education research* (pp. 427-438). Routledge.

Gliner, J. A., Morgan, G. A., & Leech, N. L. (2011). *Research methods in applied settings: An integrated approach to design and analysis*. Routledge. https://doi.org/10.4324/9780203843109

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Hartung, J., Knapp, G., & Sinha, B. K. (2011). *Statistical meta-analysis with applications*. John Wiley & Sons.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods*, *1*(1), 39-65.https://doi.org/10.1002/jrsm.5

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychological methods*, *11*(2), 193.https://doi.org/10.1037/1082-989X.11.2.193

Hussain, R., Hassali, M. H., Patel, M., & Babar, Z. (2019). Publication bias. *Encyclopedia of pharmacy practice and clinical pharmacy*.

Iacobucci, D., Popovich, D. L., Moon, S., & Román, S. (2023). How to calculate, use, and report variance explained effect size indices and not die trying. *Journal of Consumer Psychology*, *33*(1), 45-61.https://doi.org/10.1002/jcpy.1292

Jiang, Y., Ekono, M., & Skinner, C. (2017). Basic facts about low-income children: Children under 18 years. *New York: National Center for Children in Poverty, Columbia University Mailman School of Public Health*.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.https://doi.org/10.1111/jedm.12000

Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues*. Wadsworth/Thomson Learning.

Na-Nan, K., Chaiprasit, K., & Pukkeeree, P. (2018). Factor analysis-validated comprehensive employee job performance scale. *International Journal of Quality & Reliability Management*, *35*(10), 2436-2449.https://doi.org/10.1108/IJQRM-06-2017-0117

Odeyemi, J. A. (2016). A Meta–Analysis of Empirical Studies on Gender Differences in Mathematics Performance of Secondary Schools Students in Nigeria (1995-2010). Unpublished Phd. Thesis, Obafemi Awolowo University, Ile-Ife.

Ogbonnaya, U. I. (2019). The Reliability of Students' Evaluation of Teaching at Secondary School Level. *Problems of Education in the 21st Century*, *77*(1), 97-109. https://doi.org/10.33225/pec/19.77.97

Olaleye, O. C. (2015). The Development and Validation of a Scale to Measure Teacher's Disposition among Student Teachers in Osun State. An unpublished Masters Thesis. Obafemi Awolowo University, Ile-Ife.

Oswald, F. L., & McCloy, R. A. (2013). Meta-analysis and the art of the average. In *Validity Generalization* (pp. 311-338). Psychology Press.

Ovute, A. O. (2015). Meta-Analysis of Research Findings on Influence of School Location on Students' Achievement in Mathematics. *American-Eurasian Journal of Scientific Research*, *10*(1), 18-21.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage. https://doi.org/10.4135/9781412984898

Rossiter, J. R. (2010). *Measurement for the social sciences: The C-OAR-SE method and why it must replace psychometrics*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-7158-6

Sedgwick, P., & Marston, L. (2015). How to read a funnel plot in a meta-analysis. British Medical Journal, *351*.https://doi.org/10.1136/bmj.h4028

www.journal.unublitar.ac.id/jp
E-ISSN: 2598-2877, P-ISSN: 2598-5175
Vol 8 No 4, Oktober 2024

Sherif, V. (2018, March). Evaluating preexisting qualitative research data for secondary analysis. In *Forum qualitative sozialforschung/forum: Qualitative social research* 19(2).

Smith, T. R. (2013). *Evaluating the Effects of Non-Anonymity on Student Team-Member Evaluations*. Brigham Young University.

Subar, F. A. (2014). A Meta-Analytic Assessment of Empirical Differences on Standard Setting Procedures in Public Examinations. Unpublished PhD Thesis, Obafemi Awolowo University, Ile-Ife.

Trinquart, L., Dunn, A. G., & Bourgeois, F. T. (2018). Registration of published randomized trials: a systematic review and meta-analysis. *BMC Medicine*, *16*, 1-13.https://doi.org/10.1186/s12916-018-1168-6

Turner, H. M. I., & Bernard, R. M. (2006). Calculating and synthesizing effect sizes. *Contemporary issues in communication science and disorders*, *33*(Spring), 42-55.https://doi.org/10.1044/cicsd_33_S_42

Whiston, S. C. (2009). Principles and applications of assessment in counselling. *Thomson Brooks/Cole*, *2*.

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94-123.https://doi.org/10.1177/2515245919885611

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The counseling psychologist*, *34*(6), 806-838.https://doi.org/10.1177/0011000006288127

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (2002). The construction and use of psychological te'sts and measures. *Encyclopedia of life support systems*, 1-28.