SSRG International Journal of Humanities and Social Science
ISSN: 2394-2703/ https://doi.org/10.14445/23942703/IJHSS-V13I2P103

Volume 13 Issue 2, 26-32, Mar-Apr 2026

Review Article

Quality Control Strategies for Research Data Collection Instruments

Bostley Muyembe Asenahabi[1], Titus Mukisa Muhambe[2]

[1, 2]*Alupe University, Busia, Kenya.*

[1]*Corresponding Author : basenahabi@au.ac.ke*

Received: 10 February 2026 Revised: 14 March 2026 Accepted: 29 March 2026 Published: 13 April 2026

Abstract *- Quality control in data collection instruments is vital for ensuring the integrity and applicability of research findings. Poorly validated or unreliable tools can compromise measurement accuracy, weaken causal inferences, and limit generalizability. To achieve quality studies, researchers should integrate multiple forms of validity testing, such as face, content, construct, and criterion validity, alongside diverse reliability assessments such as internal consistency, test–retest, and inter-rater reliability. This ensures instruments comprehensively measure intended constructs and consistently yield stable results across contexts. At the study level, internal validity can be strengthened through randomization, control groups, standardized procedures, and elimination of confounders. External validity can be achieved through representative sampling, replication across diverse contexts, ecological relevance, and cross-validation. Together, these strategies minimize measurement error, enhance reproducibility, and advance methodological rigor. This ultimately safeguards the credibility and impact of empirical research.*

Keywords *- Quality Control, Validity, Reliability, Internal Validity, External Validity, Data Collection Instruments.*

1. Introduction

Quality control in data collection instruments continues to be a persistent challenge in empirical research. Poorly validating and having unreliable data collection tools compromises the quality, integrity, and applicability of research findings (Anastasi & Urbina, 2017). Many researchers do not adequately distinguish between the psychometric properties of instruments and research validity at the study level, which leads to measurement errors, weak causal inferences, and limited generalizability. This problem compromises the research reproducibility and rigor, particularly in quantitative studies where precision and consistency are crucial.

The core components of data collection instrument quality are validity and reliability (Karnia, 2024). Validity, which includes psychometric types like face validity, content validity, construct validity, and criterion-related validity, is the degree to which an instrument measures what it claims to measure. Conversely, reliability emphasizes consistency. A reliable data collection instrument guarantees similar outcomes when repeated under similar conditions. It can be achieved through internal consistency test, test-retest reliability, and inter-rater reliability test.

It is essential to distinguish these psychometric properties from research validity, which pertains to the study's design and inferences. Internal validity is the degree to which causal relationships between variables can be reliably established, free from confounding variables (West & Thoemmes, 2010). It assesses the accuracy of the study's causal relationships, free from confounding variables. Internal validity is achieved through randomizing the process of assigning participants to groups with an aim of balancing confounding variables, using control groups to isolate the effect of intervention, using standardized procedures, and eliminating confounders (Maxwell et al., 2015).

External validity is the degree to which the research findings can be generalized to other populations, settings, treatments, and measurement variables. It focuses on how well the research findings extend beyond the particular study environment. In order to achieve external validity, researchers use probability and random sampling techniques to select samples that closely resemble target populations in order to avoid selection bias and enhance generalizability (Hair et al., 2019), ensure replication across different contexts, ecological validity, and cross-validation.

The purpose of this study is to provide a comprehensive guide to quality control strategies for quantitative data collection instruments and clarify the distinction between psychometric validity/reliability and research validity. By addressing this gap, the study seeks to enhance

methodological rigor, minimize measurement error, and improve the reproducibility of quantitative research findings.

2. Validity

Validity is the extent to which a measurement tool accurately captures the construct it's been designed to measure. When researchers plan to conduct a study, they develop a measure that is appropriate for the study. They are expected to demonstrate that the instrument accurately captures the concept it is intended to measure. Just as arrows hitting the bullseye show precision in targeting, a valid tool consistently aligns with the "true target" of the construct under study.

Validity is a critical aspect of instrument quality. It ensures that the instrument accurately measures the intended construct. It encompasses multiple types, each addressing distinct aspects of an instrument's appropriateness and effectiveness, from superficial assessments to empirical correlations with external criteria. To establish evidence for the overall validity of a measure, researchers utilize one or more of the following types of validity.

2.1. Face Validity

Face validity is a subjective, surface-level judgment where a measurement instrument appears to assess the intended construct. It frequently relies on first impressions from experts or users without the need for empirical validation (Anastasi & Urbina, 2017). This qualitative assessment determines whether the items on the data collection instrument appear relevant and appropriate at first glance, serving as a preliminary check in psychometric development (Kline, 2016). For instance, a researcher develops a survey instrument to measure user satisfaction with a newly designed mobile application interface. To assess its face validity, the researcher sends the survey to both user experience experts and potential end-users. They are asked whether the questions clearly reflect aspects of user satisfaction (e.g., ease of navigation, clarity of icons, responsiveness). If both groups agree that the data collection instrument items appear to measure user satisfaction with the interface, the researcher can conclude that the instrument demonstrates high face validity in the computing context.

Face validity is considered a relatively weak form of validity due to its subjective nature and lack of rigorous statistical backing. Although it is a simple way of conducting a validity test, critiques argue that it does not guarantee the instrument's true measurement accuracy because appearances can deceive; a test may "appear" valid but fail to correlate with theoretical expectations or external criteria (Kline, 2016). If quantitative validation is not used to back up face validity, it can lead to misleading interpretations (Tavakol & Dennick, 2011). Nevertheless, it serves a practical role in early instrument development, enhancing participant engagement and reducing attrition by ensuring the tool seems approachable and relevant (Yusoff, 2019).

When carrying out research, face validity is often integrated with other validation types, where it acts as a gateway to more sophisticated analyses. For example, initial face validity checks precede construct validity, which is achieved through factor analysis, ensuring that items are not only theoretically sound but also user-friendly (Hair et al., 2019). Despite its weaknesses, proponents note its utility in exploratory phases, particularly in diverse cultural settings where subjective perceptions can inform adaptations (McHugh, 2012).

2.2. Content Validity

Content validity refers to the extent to which an instrument covers the entire domain of the construct being measured. It ensures that a questionnaire's items adequately represent the construct being measured by focusing on relevance and coverage. Content validity is a subjective type of validity that is accomplished through expert review panels, where they assess how well each item aligns with the intended domain. Content Validity Indices (CVI) are used to quantify the process. To implement this, researchers first assemble a panel of 5-10 subject matter experts, selected based on their knowledge level to promote objectivity and diversity. Clear evaluation criteria are then provided, guiding experts to rate the items based on relevance (typically on a 1-4 scale where 1 indicates not relevant and 4 signifies highly relevant) and coverage, assessing the extent to which the items represent key aspects of the construct. Experts evaluate the questionnaire items independently and offer ratings and qualitative feedback, which may include suggestions for revisions, deletions, or additions to enhance comprehensiveness. Feedback is collected and analyzed iteratively to refine the instrument (Yusoff, 2019).

Quantification relies on Content Validity Indices, including the Item-Level CVI (I-CVI), calculated as the proportion of experts rating an item as relevant (e.g., 3 or 4 on a 4-point scale). A threshold of ≥ 0.78 indicates strong agreement. The Scale-Level CVI (S-CVI) averages the I-CVIs across all the items, and a proportion that is rated relevant by all experts having a value ≥ 0.80 is considered to have a good overall validity. Universal Agreement (UA) tracks the percentage of items with 100% consensus, serving as a stringent check. By identifying and addressing weak items, these statistical thresholds help lower the measurement error. Following this, pilot testing confirms the questionnaire's validity in real-world applications, ensuring that it comprehensively and accurately captures the construct (Polit & Beck, 2012).

Consider a scenario where a researcher develops a questionnaire to measure employees' cybersecurity awareness. To establish content validity, the instrument is

reviewed by a panel of cybersecurity experts (e.g., IT security officers, penetration testers, compliance specialists). Item-Level CVI (I-CVI): Each expert rates the relevance of every item on a 4-point scale (1 = not relevant, 4 = highly relevant). The I-CVI is calculated as the proportion of experts rating an item as 3 or 4. Items with an I-CVI ≥ 0.78 are considered strongly valid. Scale-Level CVI (S-CVI): The overall validity of the questionnaire is assessed by averaging the-CVIs across all items or by calculating the proportion of items rated as relevant by all experts. A threshold of ≥ 0.80 indicates good overall content validity. Universal Agreement (UA): The percentage of items with 100% expert consensus is tracked as a stringent measure of content validity. Items falling below these thresholds are revised or removed to reduce measurement error. Finally, a pilot test is carried out with a sample of employees to confirm that the questionnaire is practical, understandable, and comprehensive in capturing the construct of cybersecurity awareness.

2.3. Construct Validity

Construct validity is used to evaluate data collection instruments by determining whether they accurately measure the theoretical construct they are designed to assess. It ensures that observed scores reflect the intended concept rather than being influenced by extraneous factors (Hair et al., 2019). It encompasses two primary subtypes: convergent validity, which measures the degree to which the instrument correlates positively with other established measures of the same construct, and divergent validity, which evaluates the extent to which it does not correlate with measures of unrelated constructs (Kline, 2016). To achieve construct validity, researchers employ several empirical strategies, including the use of factor analysis to examine the instrument's underlying structure and comparison with established measures, convergent validity, and divergent/discriminant validity.

2.3.1. Convergent Validity

To establish convergent validity, researchers correlate a new programming skills assessment with related tools. For example, they administer the new test alongside an established coding proficiency measure such as the HackerRank or CodeSignal assessment, then calculate Pearson's r coefficients, where values of 0.50 or higher indicate strong alignment (Bagozzi & Yi, 2012). Researchers also formulate hypotheses based on programming education theory—for instance, expecting higher correlations between the new test and students' course grades in programming modules. These hypotheses are tested statistically across diverse samples of learners to ensure generalizability. To control for confounding variables such as prior computing experience or general academic performance, partial correlations or regression analyses are employed. This combination of correlation analysis, theoretical grounding, and statistical controls provides robust evidence that the instrument truly measures the intended construct of programming skills.

2.3.2. Divergent/Discriminant Validity

Divergent/Discriminant validity is established by correlating the instrument with measures of unrelated constructs, for instance, comparing students' scores on the programming test with results from a communication skills survey or a creativity inventory, which are not theoretically linked to programming ability. If the correlations remain low ($r < 0.30$), this demonstrates discriminant power (Hair et al., 2019). It confirms that the programming assessment accurately measures the construct of programming skills by confirming that it is not unintentionally measuring unrelated traits.

2.3.3. Factor Analysis

Factor analysis plays a key role in assessing the instrument's structure: Exploratory Factor Analysis (EFA) identifies latent factors, such as ensuring personality questionnaire items load onto traits like extraversion, while Confirmatory Factor Analysis (CFA) tests hypothesized models using fit indices like RMSEA < 0.08 and CFI > 0.90 (Fabrigar et al., 2019). Implementation requires large samples ($n > 200$), evaluation of factor loadings (> 0.40), and item refinement to eliminate cross-loadings. Additionally, comparing scores with established measures involves concurrent validation, where the new instrument and a gold-standard tool are administered simultaneously, and predictive validation, assessing how well scores forecast future outcomes, such as a stress scale predicting burnout (Bagozzi & Yi, 2012). Cross-validation across split samples helps prevent overfitting and ensures robustness.

Consider a scenario where a researcher develops a test to measure undergraduate students' programming skills in Python. To establish construct validity, several empirical strategies are employed:

Factor Analysis: The test includes items on syntax knowledge, debugging, algorithm design, and problem-solving. Factor analysis is conducted to examine whether these items cluster into meaningful dimensions that reflect the underlying construct of "programming skills."

Convergent Validity: Scores from the new test are compared with students' grades in programming courses and performance on established coding challenges (e.g., HackerRank or CodeSignal). Strong correlations indicate that the instrument converges with other recognized measures of programming ability.

Discriminant/Divergent Validity: The test scores are compared with unrelated constructs, such as students' communication skills or general math anxiety. Weak

correlations demonstrate that the instrument is not measuring unrelated abilities, confirming discriminant validity.

Comparison with Established Measures: The instrument is benchmarked against standardized programming assessments or widely used curriculum rubrics to ensure alignment with recognized standards in computing education. If the factor analysis supports the hypothesized structure, the test correlates strongly with related programming measures (convergent validity), and shows weak correlations with unrelated constructs (discriminant validity), the researcher can conclude that the instrument demonstrates high construct validity in assessing programming skills.

2.4. Criterion Validity

Criterion validity refers to the extent to which an instrument's scores correlate with an external criterion. This correlation demonstrates the instrument's practical utility in measuring real-world outcomes or behaviors (Anastasi & Urbina, 2017). Criterion validity is associated with a strong correlation. It is used for evaluating data collection instruments by determining how well a measure corresponds to a standard measure or criterion. It is divided into two main types: concurrent validity, which measures the degree to which the instrument's scores align with a criterion/ gold-standard measure, and predictive validity, which evaluates the instrument's capacity to forecast future outcomes.

2.4.1. Concurrent Validity

To achieve concurrent validity, a researcher administers the data collection instrument and the criterion measure to the same respondent at the same time, after which the researcher computes correlations to quantify the relationship. For example, a marketing company creates a survey to evaluate an influencer's potential on existing social media accounts. The researcher can evaluate concurrent validity by assessing the correlation between the survey results and the account's current follower growth rate using Pearson's r or Spearman's rho coefficients. A strong correlation (e.g., $r \geq 0.70$) indicates good alignment, thus high concurrent validity (Cohen & Swerlik, 2018). This approach requires careful selection of criteria that are theoretically linked, such as comparing a depression scale to current clinical diagnoses, and controlling for variables like response bias through statistical adjustments.

2.4.2. Predictive Validity

Predictive validity involves a longitudinal design where the data collection instrument is administered to the respondents first, followed by a criterion measurement at a later point. For instance, to assess predictive validity, a company can administer a survey, then in six months compare its results to the number of new followers and sponsorship deals the account has gained (Messick, 1995). A high correlation would suggest that the account's future success can be predicted by its survey. The construct should be used to justify time intervals, and in order to preserve reliability, follow-up data collection must reduce attrition. Correlation studies form the backbone of these validations, involving bivariate or partial correlations to examine associations while accounting for confounding variables. For instance, in a study validating a health behavior survey, correlations with outcomes like exercise frequency are calculated. The effect size is interpreted using guidelines such as Cohen's (1988) thresholds for small ($r = 0.10$), medium ($r = 0.30$), and large ($r = 0.50$) effects. Regression analysis builds on this by treating the instrument as a predictor for the criterion, employing methods such as linear regression to calculate beta coefficients and R-squared values, which reveal the percentage of variance accounted for (Field, 2013). Multiple regression can include extra predictors, like demographic factors, to identify the instrument's distinct impact. To assess predictive validity, logistic regression can be used for binary outcomes, such as forecasting pass/fail rates on a certification exam, while odds ratios offer understandable effect sizes. Implementation necessitates strong sampling, including varied and representative groups to guarantee generalizability, along with statistical power analyses to identify significant effects. Researchers need to tackle possible threats, such as criterion contamination or temporal changes, by employing triangulation with different validity types.

3. Reliability

Reliability is the consistency of a data collection instrument over time and across different contexts. This implies that the data collection instrument yields consistent results across repeated applications. Just as arrows clustered tightly together on a target show uniformity, a reliable tool yields comparable outcomes under identical conditions. Reliability of data collection instruments can be established through different forms: internal consistency, test-retest reliability, and inter-rater reliability.

3.1. Internal Consistency

Internal consistency refers to the extent to which items within a data collection instrument are correlated, demonstrating that they collectively measure the same underlying construct. It denotes the degree to which the items within a data collection instrument are interrelated and measure the same underlying construct. This relationship reflects the homogeneity and coherence of the instrument's responses. (Kline, 2016). Internal consistency measures how well individual items relate to each other and collectively assess the same underlying construct. This ensures uniformity and coherence in the responses generated by the instrument. Internal consistency is essential because it indicates that the items in a data collection instrument are measuring the same concept without excessive noise or unrelated variance. This, in turn, supports the instrument's internal structure and enhances its overall dependability.

To achieve and assess internal consistency, researchers mainly employ two established methods: Cronbach's alpha and split-half reliability.

3.1.1. Cronbach's Alpha

Cronbach's alpha quantifies the average inter-item correlations within a scale. It provides an estimate of reliability by examining the covariance among items relative to the total variance (Tavakol & Dennick, 2011). Cronbach's alpha values typically range between 0 and 1. Coefficient values above 0.70 are generally considered acceptable in most research contexts and indicate good internal consistency. Values between 0.60 and 0.70 may be acceptable for exploratory scales. Scores below 0.60 suggest that revisions to the instrument are necessary to improve reliability (Kline, 2016).

To compute Cronbach's alpha, researchers collect data from a sample of respondents and calculate the mean inter-item correlation using statistical software like SPSS or R for efficiency. While performing this process, researchers should examine item-total correlations to identify and potentially remove poorly performing items that lower the alpha value.

3.1.2. Split-Half Reliability

Split-half reliability works by dividing the items of a data collection instrument into two equivalent halves. These halves can either be generated randomly or by using odd-even item splits. The scores from each half are then correlated to assess the degree of consistency. The Spearman-Brown prophecy formula is used to assess the reliability of the complete scale (Tavakol & Dennick, 2011). This method evaluates consistency by checking whether the two halves of the instrument produce similar results. The correlation coefficients are then adjusted for the length of the test, allowing researchers to predict the reliability of the entire instrument. For instance, if the halves correlate at $r = 0.80$, the formula $r\_sb = (2r) / (1 + r)$ yields an estimated reliability of approximately 0.89. To ensure stable reliability estimates, the application of split-half reliability requires a sufficiently large sample, usually exceeding 100. Robustness can be improved by dividing the items using several techniques, after which the obtained coefficients are averaged. Researchers should ensure item equivalence in the splits to avoid bias, and this technique is particularly useful for instruments without established norms.

Consider a study where a researcher designs a survey to evaluate mobile phone usability features, focusing on ease of navigation, responsiveness, clarity of icons, and accessibility. After collecting initial responses, the researcher calculates internal consistency metrics such as Cronbach's alpha to determine how well the items collectively measure usability. If certain items show weak correlations (values ≥ 0.6) with the overall scale, they are carefully reviewed, revised, or eliminated to improve coherence. Through this iterative refinement process, the instrument evolves into a more reliable tool. A subsequent round of testing demonstrates stronger internal consistency, confirming that the survey consistently captures the construct of mobile phone usability. This approach ensures that the instrument is both statistically sound and practically effective in assessing user perceptions.

3.2. Test-Retest Reliability

Test-retest reliability is the degree to which a data collection instrument yields consistent results when administered to the same participants on two separate occasions. It reflects the temporal stability of the measurements and shows that the instrument is free from transient errors. It ensures that observed changes in scores are the result of genuine variations rather than instability in the measurement process.

To achieve test-retest reliability, the data collection instrument must be administered twice to the same group of participants at different time points. The correlation between the two sets of scores is then analyzed to determine consistency. The first step in this process is to select an appropriate sample. Ideally, the sample should consist of 50 to 200 participants who are representative of the target population. Representativeness is ensured by considering factors such as age, gender, and cultural background.

This helps minimize confounding influences and supports the generalizability of the findings. (Field, 2013). The time interval between administrations should be carefully chosen based on the construct being measured; for stable traits like personality, intervals of 2-4 weeks are common to avoid memory effects, whereas for dynamic constructs like mood, shorter intervals (e.g., 1-2 days) may be used, but longer ones (e.g., 6-12 months) can assess long-term stability (Kline, 2016). Participant retention is essential and can be achieved through follow-up strategies, such as reminders or incentives, to reduce attrition, which could bias results.

After gathering the data, the relationship between the two sets of scores is determined using statistical methods such as Pearson's r for continuous data or Intraclass Correlation Coefficients (ICC) for a stronger evaluation of consistency. ICC values above 0.70 typically indicate good reliability (Cohen & Swerlik, 2018). For example, when validating a stress inventory, scores from Time 1 and Time 2 could correlate at $r = 0.85$, indicating high stability. Interpretation involves examining the correlation's strength, where values closer to 1.0 denote excellent consistency, and lower values (e.g., below 0.60) may indicate issues such as measurement error or true changes in the construct. Paired sample t-test can also be performed to check for systematic differences between administrations to ensure that there is no significant shift due to practice effects or external events.

3.3. Inter-Rater Reliability

Inter-rater reliability indicates the degree of agreement among different raters or observers when assessing the same data using a standardized data collection instrument. It reflects the consistency of their judgments and demonstrates the objectivity of the measurement process. (McHugh, 2012). This form of reliability is particularly vital for subjective or observational data collection instruments, such as rating scales or checklists, where human judgment plays a role in scoring.

To achieve inter-rater reliability, researchers employ systematic approaches that emphasize rater preparation and statistical evaluation of agreement. It is essential to train the raters so that they have a shared understanding of the instrument's criteria, scoring guidelines, and potential biases. The training process often involves workshops, practice sessions with sample data, and calibration exercises to align interpretations (McHugh, 2012). For instance, in a study using a behavioral observation scale, the observers could analyze video recordings together, address discrepancies, and refine their application of categories until consensus is reached. This training phase should incorporate clear definitions of constructs, examples of high and low scores, and strategies that can be used to minimize subjective influences, such as anchoring biases.

After training, the instrument is used on the same set of data by multiple raters independently, followed by the calculation of agreement metrics to measure consistency. Frequent statistical methods include Cohen's kappa for categorical data, which accounts for agreement occurring by chance, with values above 0.60 generally indicating substantial reliability (Hallgren, 2012). Intraclass Correlation Coefficients (ICC) are preferred for continuous or ordinal scales since they measure the variance due to actual agreement, with ICC values of 0.75 or higher suggesting excellent consistency (Koo & Li, 2016). Researchers compute these using software like SPSS or R, analyzing pairwise or multi-rater correlations to identify patterns.

4. Recommendations

To ensure a researcher ends up with a quality study, this study recommends that researchers integrate multiple forms of validity testing alongside diverse reliability assessments to ensure that instruments comprehensively measure intended constructs and consistently produce stable results across contexts. Besides, assessing both internal and external validity is essential for ensuring reproducibility and advancing methodological rigor in research.

5. Conclusion

Ensuring quality in data collection instruments is fundamental to credible research. By integrating multiple forms of validity testing and diverse reliability assessments, researchers can develop tools that accurately measure intended constructs and consistently yield stable results. Distinguishing psychometric validity from research validity further strengthens methodological rigor. Assessing both internal and external validity enhances reproducibility, minimizes measurement error, and safeguards the integrity and generalizability of quantitative research findings.

Conflicts of Interest

There is no conflict of interest regarding the publication of this paper.

Funding Statement

The publication of this article has been funded by the authors.

References

[1] Anne Anastasi, and Susana Urbina, *Psychological Testing*, 8th Ed., Pearson, 2017. [Google Scholar] [Publisher Link]

[2] Richard P. Bagozzi, and Youjae Yi, "On the Evaluation of Structural Equation Models," *Journal of the Academy of Marketing Science*, vol. 40, no. 1, pp. 34-50, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ronald Jay Cohen, and Mark E. Swerdlik, *Psychological Testing and Assessment*, 8th Ed., McGraw-Hill, 2018. [Google Scholar] [Publisher Link]

[4] Leandre R. Fabrigar et al., "Evaluating the Use of Exploratory Factor Analysis in Psychological Research," *Psychological Methods*, vol. 4, no. 3, pp. 272-299, 2019. [Google Scholar] [Publisher Link]

[5] Andy Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th Ed., Sage Publications, 2013. [Google Scholar] [Publisher Link]

[6] Joseph F. Hair et al., *Multivariate Data Analysis*, 8th Ed., Cengage, 2019. [Google Scholar] [Publisher Link]

[7] Kevin A. Hallgren, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," *Tutorials in Quantitative Methods for Psychology*, vol. 8, no. 1, pp. 23-34, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[8] Richard Karnia, "Importance of Reliability and Validity in Research," *Psychology and Behavioral Sciences*, vol. 13, no. 6, pp. 137-141, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rex B. Kline, *Principles and Practice of Structural Equation Modeling*, 4th Ed., Guilford Press, 2016. [Google Scholar] [Publisher Link]

[10] Terry K. Koo, and Mae Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155-163, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[11] S.E. Maxwell et al., “Is Psychology Suffering from a Replication Crisis? What Does "Failure to Replicate" Really Mean?,” *American Psychologist*, vol. 70, no. 6, pp. 487-498, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[12] Mary L. McHugh, “Interrater Reliability: The Kappa Statistic,” *Biochemia Medica*, vol. 22, no. 3, pp. 276-282, 2012. [Google Scholar] [Publisher Link]

[13] Samuel Messick, “Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning,” *American Psychologist*, vol. 50, no. 9, pp. 741-749, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[14] Denise F. Polit, and Cheryl Tatano Beck, *Nursing Research: Generating and Assessing Evidence for Nursing Practice*, 9th Ed., Lippincott Williams & Wilkins, 2012. [Google Scholar] [Publisher Link]

[15] Mohsen Tavakol, and Reg Dennick, “Making Sense of Cronbach's Alpha,” *International Journal of Medical Education*, vol. 2, pp. 53-55, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[16] Stephen G. West, and Felix Thoemmes, “Campbell's and Rubin's Perspectives on Causal Inference, *Psychological Methods*, vol. 15, no. 1, pp. 18-37, 2010. [Google Scholar] [Publisher Link]

[17] Muhamad Saiful Bahri Yusoff, “ABC of Content Validation and Content Validity Index Calculation,” *Education in Medicine Journal*, vol. 11, no. 2, pp. 49-54, 2019. [CrossRef] [Google Scholar] [Publisher Link]