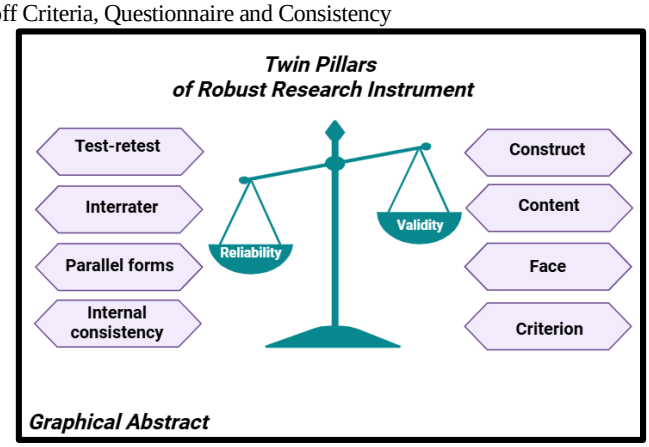
MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2225-2230

Validity and Reliability: The Twin Pillars of Robust Research Instrument

Rishika Awasthi[1*],
Assistant Professor, Department of Management,
GL Bajaj Institute of Management, Greater Noida, UP, India.
ORCID- 0009-0000-6124-2802, Email ID: rishikaawasthi95.ra@gmail.com
Rashmi Rakesh[2],
Assistant professor, School of Business Management,
Maharishi University of Information Technology, Lucknow, UP, India.
Kshitij Shukla[3],
Department Of Commerce,
Integral University, Lucknow, UP, India.
Priyanka Bajpai[4],
Assistant Professor, Department of Business Management,
Integral Business School, Integral University, Lucknow, UP, India.
Rachna Vats[5],
Assistant Professor, Department of Management,
GL Bajaj Institute of Management, Greater Noida, UP, India.

Abstract
Reliability and Validity are crucial for research accuracy and credibility, especially in quantitative studies. Reliability refers to the consistency of a method, ensuring consistent results under similar conditions, while validity measures whether the method accurately measures intended results. Assessing reliability through methods like test-retest, inter-rater reliability, or internal consistency, and validity through content, construct, and criterion validity, ensures trustworthy findings. However, these measures are often overlooked by health and social science researchers in developing countries due to a lack of knowledge. This article explores reliability and validity tests providing insights for improving research quality.
Keywords: Reliability, validity, Cutoff Criteria, Questionnaire and Consistency



INTRODUCTION

Reliability and Validity are crucial attributes for evaluating measurement instruments in research. Validity measures the accuracy and consistency of measurements, while reliability reflects the data's dependability and ability to minimize random errors. Despite their importance, health, and social science researchers in developing countries often overlook these evaluations due to a lack of understanding of these concepts. This review article provides an in-depth examination of the validity and reliability of research instruments, focusing on questionnaires. It discusses different types of validity and reliability tests, their methods, and the scientific principles underpinning these analyses.

Reliability and Validity are crucial for transparency and minimizing researcher bias in qualitative research (Singh, 2014). Assessing reliability and validity requires evaluating methodologies used during data collection (Saunders et al., 2009). These concepts are essential for interpreting psychometric instruments like symptom scales, questionnaires, educational tests, and observer ratings (Cook & Beckman, 2006). Validity and reliability enhance the precision and rigor of assessment and evaluation processes (Tavakol & Dennick, 2011). Without evaluating these attributes, accounting for measurement errors' impact on theoretical constructs becomes difficult (Forza, 2002). Using diverse and appropriate data collection methods improves the validity and reliability of data, ensuring the accuracy of findings.

Researchers often overlook the importance of measuring reliability and scale validity in research (Thompson, 2003). Measurement involves assigning numerical values to observations to quantify phenomena, including the construction of variables and the development of instruments or tests (Kimberlin & Winterstein, 2008). Employing robust mechanisms enhances scientific rigor, enabling accurate measurement of variables and ensuring the credibility of findings. Errors are most likely in measuring scale variables, making the development of reliable and valid scales essential for producing high-quality research (Shekharan & Bougie, 2010). Measurement errors hinder the ability to detect significant results and compromise the utility of scores in generating reliable findings. Establishing reliability and validity in research ensures robust, replicable, and accurate data. Reliability is a term derived from the words "rely on" and "ability," referring to the dependability of a measurement tool (Wahyudi, 2020). It signifies the consistency of measurement results when repeated. Reliable measurements are those that consistently produce dependable data (Arikunto, 2010). Tests are considered reliable if they show a strong correlation between observed scores and actual scores (Arifin, 2017). The instrument itself influences the validity and reliability of a measurement instrument, the individuals using it, and the characteristics of the subjects being measured (Sugiyono, 2014). Whereas, validity is the degree to which a measuring instrument accurately fulfills its intended function, ensuring it accurately reflects the intended variable (Wahyudi, 2020). It is defined as the test's ability to accurately measure what it is designed to measure, ensuring the instrument accurately aligns with the actual conditions of the variable being measured without deviation (Arifin, 2017; Yusup, 2018).

Reliability and Validity are essential attributes of a measuring instrument. While both are interconnected, they are distinct attributes. A reliable instrument can be valid without being valid, but it is generally reliable. However, reliability alone doesn't guarantee validity. Researchers must evaluate both validity and reliability to ensure the accuracy and credibility of research findings. Misinterpreting results may lead to misleading conclusions.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2225-2230

Figure 1: Concept of Reliability and Validity

REQUISITE OF RELIABILITY AND VALIDITY OF INSTRUMENT (QUESTIONNAIRE)

Questionnaires are versatile tools used in health and social science research, used in various survey methods such as postal, electronic, face-to-face (F2F), and telephone (Bolarinwa, 2015). Postal and electronic questionnaires are self-completion, completed independently by respondents.F2F and telephone questionnaires are administered by interviewers, who pose standardized questions and record responses (Bolarinwa, 2015). Questionnaires can be adapted from validated instruments or developed to measure specific attributes. To ensure accuracy and consistency, it is crucial to test the validity and reliability of these tools (Bolarinwa, 2015; Kember et al., 2008; Wong et al., 2012).

RELIABILITY AND RELIABILITY COEFFICIENT

Reliability is the consistency, precision, repeatability, and trustworthiness of research findings, ensuring they are free from bias (Bordens, & Abbott, 2014; Mohajan, 2017). It is crucial in qualitative research for validity and reflects the accuracy of an assessment tool (Mohajan, 2017). In quantitative research, reliability reflects the stability and reproducibility of results obtained under identical conditions but in varying circumstances, such as across different researchers or projects (Mohajan, 2017).

The reliability coefficient is a statistical measure that quantifies the stability of measurements across different instances for the same items (Cohen et al., 2022). It can assess various forms of reliability, ranging from 0 to 1. For high-stakes measurements, a reliability coefficient exceeding 0.9 is recommended, while a coefficient greater than 0.8 is generally considered high. Accurately measuring reliability enhances researchers' confidence in adopting new testing methodologies, as higher reliability reflects greater accuracy in data, increasing the likelihood of making informed research decisions (Mohajan, 2017).

MEASUREMENT TOOLS OF RELIABILITY

Test-retest reliability is the consistency of a measurement administered to a group of individuals at two different points in time, with a correlation coefficient of +0.80 or higher. A high-reliability score indicates a correlation of +0.80 or higher, and a specific time interval between the two measurements is crucial (Price, et al., 2017). A low correlation may indicate too much time has passed, maturation has occurred, or errors in the measurement process exist (Mohajan, 2017). Parallel-forms reliability assesses the consistency between two different assessment tools given to the same group of individuals. The scores from the two versions are correlated to determine the relationship between them. Items on a parallel form are designed to be equivalent to those on the original form, assessing the same knowledge and skills but using different questions to avoid recall bias from a prior assessment (Mohajan, 2017). Split-half reliability evaluates internal consistency by comparing one-half of the results against the other half of an assessment (Mohajan, 2017). Internal reliability refers to the consistency of responses across different items measuring the same underlying construct (Ahmed & Ishtiaq, 2021). High internal consistency indicates that different sections of a test yield consistent results when measuring the same construct. A t-test can be employed to identify significant statistical differences between the scores when comparing reliability data across different factors (Knodt et al., 2023).

TYPES OF RELIABILITY

Reliability is mainly divided into two types: i) stability, and ii) internal consistency reliability.

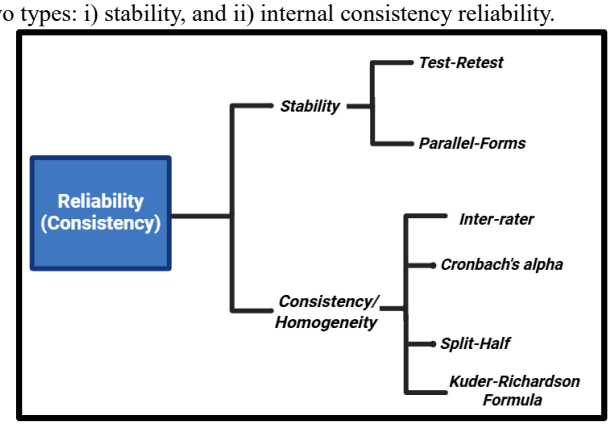


Figure 2: Types of Reliability

STABILITY (MEASURES ERROR BECAUSE OF DIFFERENCES IN TEST FORMS)

The stability of a measure refers to its consistency over time, even under uncontrolled conditions or variations in respondents. It indicates an individual's score's likelihood of remaining unchanged between two administrations (Allen & Yen, 1979). A perfectly stable measure would yield identical scores on each administration.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2225-2230

METHODS FOR TESTING STABILITY

1. **Test-retest reliability** is crucial for establishing the utility of an assessment tool and predicting generalizable outcomes (Schaaf et al., 2024). Parameter identifiability is a key prerequisite for ensuring test-retest reliability, as it guarantees stability over time (Liu et al., 2024). Factors influencing test-retest reliability include participant behavior stability, cognitive and mood changes over time, and the accuracy of the original measurement (Schaaf et al., 2024).
2. **Parallel-forms reliability** assesses the consistency of results obtained from two different versions of an assessment tool administered to the same group of individuals. The correlation between the scores from these two versions helps evaluate the reliability of the measure (DeVellis, 2006). For example, employee satisfaction levels can be assessed using multiple methods, such as questionnaires, in-depth interviews, and focus groups. If the results from these methods are highly correlated, the measures are considered reliable (Yarnold, 2014).

INTERNAL CONSISTENCY RELIABILITY / HOMOGENEITY (MEASURES ERROR BECAUSE OF IDIOSYNCRASIES OF THE TEST ITEMS)

Internal consistency reliability evaluates the extent to which different items in a test that measure the same construct yield similar results (DeVellis, 2006). This form of reliability can be assessed in a single testing instance, circumventing issues of repeated testing associated with other reliability measures (Allen & Yen, 1979). Internal consistency can be represented in different formats (Cortina, 1993).

1. **Inter-rater reliability** Inter-rater reliability refers to the degree to which different raters consistently assess the same information (Keyton et al., 2004). It ensures that ratings obtained with an instrument are equivalent when used by different observers.
2. **Cronbach's alpha (α)** is the most common measure of internal consistency, which ranges from 0 to 1 (Tavakol & Dennick 2011), with values above 0.7 being generally deemed acceptable, above 0.8 being good (Nunnally & Bernstein, 1994), and above 0.9 being exceptional (Cronbach, 1951).
3. **Split-half reliability** measures internal consistency by comparing one-half of a test's items against the other half (Thanasegaran, 2009). It is particularly useful when the test is lengthy and requires only one administration (Chakrabartty, 2013).
4. **Kuder-Richardson Formula 20 (K-R 20) and Kuder-Richardson Formula 21 (K-R 21)** measure the consistency of subjects' responses to dichotomous scoring tests. They compare all test items, unlike split-half methods. K-R 20 is more general, assuming equal difficulty for all questions, while K-R 21 assumes equal difficulty. K-R coefficients can also be seen as the average of all possible split-half reliability estimates using the Rulon formula (Kuder & Richardson, 1937; Oluwatayo, 2012).

VALIDITY AND VALIDATION

Validity is the degree to which a research instrument accurately measures what it was intended to measure (Bordens, & Abbott, 2014). In experimental design, the objective is to establish robust test-retest reliability and internal consistency, ensuring that the results genuinely reflect the intended construct (Price, et al., 2017). Validation involves analyzing collected data and assessing its validity (Cohen et al., 2022). In qualitative research, validity is determined by the rigorous application of the scientific method, focusing on the trustworthiness, utility, and dependability of results. In quantitative research, validity pertains to how accurately a measurement device or test captures the intended variable (Mohajan, 2017).

Research validity consists of two components: internal and external validity. Internal validity assesses the study's legitimacy based on the design, including sample selection, data collection, and analysis procedures (Mohajan, 2017). External validity evaluates whether the findings are generalizable to other populations or settings (Mohajan, 2017). Without external validity, researchers may fail to establish cause-and-effect relationships and undermine the ability to make broader claims about the findings (Esterling et al., 2023). To enhance validity, researchers should carefully plan their studies, incorporating quality control measures such as effective recruitment strategies, accurate data collection methods, robust data analysis, appropriate sample sizes, and criteria reflecting real-world scenarios (Patino & Ferreira, 2018).

A test must be reliable and have high internal validity. It should be developed using sound measurement principles and standardization, which involves clear and explicit methods for administering tasks. Key elements of a standardized testing environment include a quiet, distraction-free setting, scripted instructions, and necessary tools or stimuli. Standardized tests provide normative data, allowing an individual's performance to be compared to a representative sample from the intended test population (Institute of Medicine of the National Academies. 2015). These norms should reflect a broad and diverse sample, ensuring equal opportunity for each individual. Applying a test to individuals not part of the intended population or not included in the norm group can lead to inaccurate scores and misinterpretation (Institute of Medicine of the National Academies. 2015).

MEASUREMENT TOOLS OF VALIDITY

Test developers use various methods to ensure the validity of assessments, including face validity, content validity, criterion-related validity, and construct validity (Zafrullah et al., 2024). Face validity is a subjective evaluation by a subject matter expert to determine if an instrument appears relevant and appropriate (Price, et al., 2017; Setia, 2017). Content validity involves a similar assessment, ensuring that a questionnaire adequately measures the intended concepts (Setia, 2017). Criterion-related validity is the extent to which a measure aligns with an established standard or specific outcomes (Setia, 2017). This form of validity is often used to predict future performance by examining the correlation between scale scores and a measurable criterion (Mohajan, 2017). Construct validity focuses on evaluating how well items within a questionnaire relate to the underlying theoretical constructs being measured. The variables measured are not directly observable and have been developed to explain specific behaviors. To assess construct validity, it is crucial to show that individuals who score highly or low on the measure exhibit behaviors consistent with the predictions of the underlying theory (Bordens, & Abbott, 2014).

TYPES OF VALIDITY

The validity of a questionnaire can be determined through translational or representational validity, which involves a panel of experts evaluating the theoretical constructs underlying the measure as shown in figure 3. This process includes two subtypes: face validity and content validity (Bhattacherjee, 2012). Alternatively, questionnaire validity can be examined through a field test, which explores how well a given measure correlates with external criteria based on empirical constructs. This form of validity includes criterion-related validity and construct validity, with some scholars arguing that the two are distinct constructs (Bhattacherjee, 2012; Engel & Schutt, 2016). Predictive and concurrent validity are subtypes of criterion-related validity, while convergence validity, discriminant validity, known-group validity, and factorial validity are subtypes of construct validity. Hypothesis testing validity is also considered a form of construct validity (Wells & Wollack, 2003).

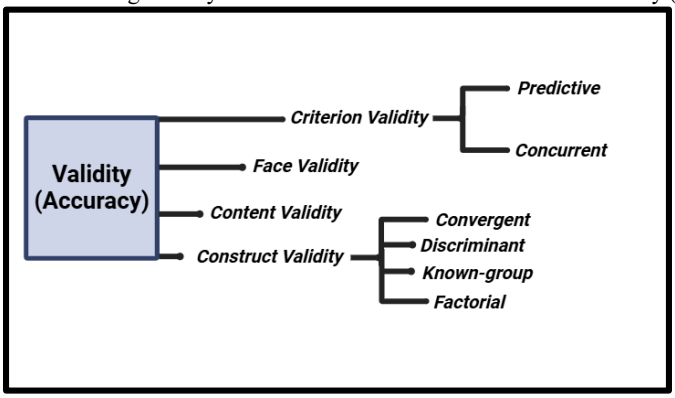


Figure 3: Types of Validity

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2225-2230

Face validity is a subjective evaluation of a questionnaire or instrument by an expert in the research subject, assessing whether each item aligns with the conceptual domain of the concept being measured (Miller, 2012; Bölenius et al., 2012). Although often considered a casual and informal form of validity (Engel & Schutt, 2016), it remains widely used, especially in developing countries. Its simplicity and ease of implementation contribute to its popularity (Sangoseni et al., 2013).

Content validity refers to the extent to which an instrument comprehensively measures the construct intended to assess (Miller, 2012; Sangoseni et al., 2013; DeVon et al., 2007; Polit & Beck, 2006). For example, a researcher may use content validity to assess employees' attitudes toward a training program on hazard prevention. Experts review the questionnaire items for clarity, readability, and comprehensiveness, reaching a consensus on the inclusion of items (Sangoseni et al., 2013; DeVon et al., 2007; Polit & Beck, 2006).

Content validity indices (CVI) are often calculated at the item level (I CVI) and scale level (S CVI), with the S CVI reflecting the level of agreement among raters (Sangoseni et al., 2013). Readability assessments using formulas like the Fog Index, Flesch Reading Ease, Flesch-Kincaid readability formula, and Gunning Fog Index are commonly used to evaluate content validity (Miller, 2012; Wells & Wollack, 2003). However, content validity has a key limitation due to its subjective nature, similar to face validity. Researchers may combine multiple validity forms to strengthen the overall validity of a questionnaire (Sangoseni et al., 2013; Böleniu et al., 2012; Anderson et al., 2002; Mackison et al 2010).

Criterion-related validity is the degree to which test scores align with an external criterion (Drost, 2011; Liang et al., 2014), indicating how well a questionnaire's findings align with another established instrument or predictor (Wong et al., 2012; Liang et al., 2014). However, it can be complicated by the unavailability or measurable nature of the relevant predictor.

Convergent validity occurs when different measures of the same concept produce similar results, indicating consistent assessment. This can be achieved by using diverse methods to assess the same construct, such as comparing self-reported data with observational data (Wells & Wollack, 2003; Schimmack, 2010; Anderson & Sellbom, 2015; Erdvik et al., 2015; DeVellis & Thorpe, 2021). For example, a researcher could use meters on respondents' televisions to track their time spent watching health-related programs, and then compare this with survey responses. Alternatively, an observer could record respondents' TV viewing habits directly, and then compare these results with questionnaire responses. Convergent validity is demonstrated when two different measurement methods yield comparable results, thereby reinforcing the reliability of the studied concept.

Discriminant validity is the ability to distinguish one concept from other closely related concepts (Wells & Wollack, 2003; Anderson & Sellbom, 2015; DeVellis & Thorpe, 2021). For example, comparing TV health program exposure to entertainment programs would not strongly correlate, indicating that the two constructs are separate and distinct.

Known-group validity involves comparing groups with different attributes related to the study construct, with one group known to exhibit the attribute and the other not yet established (Engel & Schutt, 2016; Deniz, & Alsaffar, 2013). The assumption is that the group with the relevant attribute will demonstrate higher scores on the measured construct, while the group without the attribute will score lower (Engel & Schutt, 2016; Deniz, & Alsaffar, 2013; Singh, et al., 2011; DeVellis & Thorpe, 2021). For example, a study examining depression would expect higher scores for individuals with a clinical diagnosis of depression. Singh et al., (2011) demonstrated known-group validity by comparing cognitive interview responses among schoolchildren across six European countries.

Factorial validity is an empirical method used to validate the components of a construct using factor analysis (Engel & Schutt, 2016; Douglas et al., 2012; Motl et al., 2000; Dhillon et al., 2014; Anastasiadou, 2011). It is particularly useful when a construct encompasses multiple dimensions, each representing a distinct domain of a broader attribute. In factorial validity analysis, items measuring a specific dimension of the construct should be more strongly correlated with one another than with items measuring other dimensions ((Engel & Schutt, 2016; Douglas et al., 2012; Motl et al., 2000; Dhillon et al., 2014; Anastasiadou, 2011). For example, in the health-related Quality of Life questionnaire (SF-36v2), items assessing the "social function" dimension should show a stronger correlation with each other than those measuring the "mental health" dimension.

Hypothesis testing validity provides evidence that a research hypothesis regarding the relationship between measured variables, derived from a theory, is supported (Wells & Wollack, 2003; Parsian & Dunning, 2009). For example, if data supports the hypothesis positing a positive correlation between physical aggression and televised violence, it strengthens the construct validity of both measures, affirming that theoretical concepts are accurately measured and examined through hypothesis testing.

CUTOFF CRITERIA OF RELIABILITY AND VALIDITY TEST

Reliability and validity are crucial in assessing the quality of measurement tools and ensuring accurate research outcomes. Reliability refers to the consistency of a measurement instrument, producing stable and repeatable results over time or across different contexts. Common methods for assessing reliability include test-retest reliability, internal consistency measures like Cronbach's alpha, and inter-rater reliability. Validity evaluates whether an instrument measures what it is intended to measure, including content validity, construct validity, and criterion validity as shown in Table 1.

Table 1: Cutoff Criteria of Reliability and Validity Test

Test Name	Rationale of Test	Statistical Technique	Recommended Criteria	Range	Reference
Reliability					
Test-Retest	Assess Stability Over Time	Correlation coefficient	$r > 0.70$ for acceptable stability	0 to 1	(Nunnally & Bernstein 1994)
Split-Half	Evaluate Internal Consistency	Spearman-Brown formula	$r > 0.70$ for acceptable reliability	0 to 1	(Kline, 2000)
Kr-20	Internal Consistency for Dichotomous Items	Kuder-Richardson Formula 20	KR-2 $0 > 0.70$	0 to 1	(Kuder & Richardson, 1937)
Cronbach's Alpha	Internal Consistency for Scale Items	Cronbach's alpha	$\alpha > 0.70$	0 to 1	(Cronbach, 1951)
Parallel Or Alternate Forms Reliability	Equivalence Between Test Versions	Correlation coefficient	$r > 0.80$ for equivalence	0 to 1	(Anastasi & Urbina, 1997)
Inter-Rated Reliability	Agreement among Raters	Cohen's Kappa or ICC	kappa > 0.60 or ICC > 0.75	0 to 1	(Fleiss et al., 1981; Shrout & Fleiss, 1979; Wu et al., 2025)
Validity					
Content Validity	Assess Item Relevance to the Content Domain	Expert judgment, Content Validity Index (CVI)	CVI > 0.80	0 to 1	(Lynn, 1986)
Concurrent Validity	Correlation with Current Measures	Correlation coefficient	$r > 0.70$	0 to 1	(Anastasi & Urbina, 1997)
Predictive Validity	Ability to Predict Future Outcomes	Regression analysis	Statistically significant model	N/A	(Cronbach & Meehl, 1955)
Convergent Validity	Correlation With Related Constructs	Correlation coefficient	$r > 0.50$	0 to 1	(Campbell & Fiske, 1959)
Divergent Validity	Lack Of Correlation with Unrelated Constructs	Correlation coefficient	$r < 0.30$	0 to 1	(Campbell & Fiske, 1959)
Contrasted Groups	Differentiate Between Groups	t-test or ANOVA	Statistically significant result	$p < 0.05$	(Cohen, 2013)
Factor Analysis	Identify Underlying Constructs	Exploratory or Confirmatory Factor Analysis	Factor loadings > 0.40	0 to 1	(Tabachnick & Fidell, 2007)

THREATS TO VALIDITY AND RELIABILITY

The validity and reliability of research findings can be threatened by various factors, including errors due to researcher oversight, participant behavior, contextual influences, or methodological shortcomings (Lillis, 2006). Researchers must be diligent in identifying and addressing potential sources of error during the design and execution of their studies. Errors that affect reliability are typically random, while those that impact validity are often systematic or constant.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2225-2230

Threats to reliability include inconsistent procedures, such as unclear or non-standardized instructions, incomplete alternatives, improper question sequencing, ambiguous wording in instruments, lengthy or difficult-to-read questionnaires, or overly extended interview durations (Kerlinger, 1964; Fink & Kosecoff, 1985). These factors can compromise the consistency and repeatability of measurements.

Internal validity may be compromised at various points in the research process, particularly during data collection, analysis, or interpretation (Tashakkori & Teddlie, 1998; Onwuegbuzie, 2003). Key threats to internal validity include instrumentation errors, order bias, and researcher bias in applying research techniques. Instrumentation issues arise when measurement tools fail to produce consistent or accurate scores, order bias influences the results in a way that cannot be distinguished from the effects of the interventions themselves, and researcher bias represents a personal preference for certain techniques.

External validity is threatened when there are biases or limitations in the sample population or the settings under which the study is conducted, impacting the generalizability of the results (Hahs-Vaughn & Lomax, 2020). An example of a measurement error is the case of a clock that consistently runs five minutes fast, which is not valid as its time does not align with the actual time standard.

CONCLUSION

This article discusses the importance of questionnaires in social and health science research, emphasizing the need for valid and reliable tests. It discusses various methods for evaluating questionnaires, focusing on developing countries. The article also highlights common research errors and the fact that a reliable instrument does not guarantee validity. It also addresses potential threats to reliability and validity that researchers may face while conducting rigorous research. The article provides both theoretical and technical interpretations of these concepts.

REFERENCES

Ahmed, I., & Ishtiaq, S. (2021). Reliability and validity: Importance in Medical Research. *JPMA. The Journal of the Pakistan Medical Association*, *71*(10), 2401-2406.

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory (Brooks/Cole Publishing Company, Monterey, CA).

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall/Pearson Education.

Anastasiadou, S. D. (2011). Reliability and validity testing of a new scale for measuring attitudes toward learning statistics with technology. *Acta Didactica Napocensia*, *4*(1), 1-10.

Anderson, A. S., Bell, A., Adamson, A., & Moynihan, P. (2002). A questionnaire assessment of nutrition knowledge–validity and reliability issues. *Public health nutrition*, *5*(3), 497-503.

Anderson, J. L., & Sellbom, M. (2015). Construct validity of the DSM–5 section III personality trait profile for borderline personality disorder. *Journal of personality assessment*, *97*(5), 478-486.

Arifin, Z. (2017). Kriteria instrumen dalam suatu penelitian. *Jurnal Theorems (the original research of mathematics)*, *2*(1), 28-36.

Arikunto, S. (2010). Prosedur penelitian: Suatu pendekatan praktik. Jakarta: Rineka Cipta.

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. University of South Florida.

Bolarinwa, O. A. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian postgraduate medical journal*, *22*(4), 195-201.

Bölenius, K., Brulin, C., Grankvist, K., Lindkvist, M., & Söderberg, J. (2012). A content validated questionnaire for assessment of self-reported venous blood sampling practices. *BMC research notes*, *5*, 1-6.

Bordens, K., & Abbott, B. B. (2014). *Ebook: Research Design and Methods: A Process Approach*. McGraw Hill.

Bougie, R., & Sekaran, U. (2019). *Research methods for business: A skill building approach*. john wiley & sons.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Chakrabartty, S. N. (2013). Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education*, *3*(1), 1-8.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Cohen, R. J., Tobin, R. M., & Schneider, W. J. (2022). *Psychological testing and assessment : an introduction to tests and measurement* (Tenth Edition). McGraw Hill LLC.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, *119*(2), 166-e7.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Deniz, M. S., & Alsaffar, A. A. (2013). Assessing the validity and reliability of a questionnaire on dietary fibre-related knowledge in a Turkish student population. *Journal of health, population, and nutrition*, *31*(4), 497.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Devillis, R.E., "Scale Development: Theory and Application," Applied Social Science Research Method Series, Vol. 26 (Newbury Park: Sage Publishers Inc., 2006).

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship*, *39*(2), 155-164.

Dhillon, H. K., Zaini, M. Z. A., Quek, K. F., Singh, H. J., Kaur, G., & Rusli, B. N. (2014). Exploratory and confirmatory factor analyses for testing validity and reliability of the malay language questionnaire for urinary incontinence diagnosis (QUID). *Open Journal of Preventive Medicine*, *4*(11), 844.

Douglas, H., Bore, M., & Munro, D. (2012). Construct validity of a two-factor model of psychopathy. *Psychology*, *3*(03), 243.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and perspectives*, *38*(1), 105-123.

Engel, R. J., & Schutt, R. K. (2016). *The practice of research in social work*. Sage Publications.

Erdvik, I. B., Øverby, N. C., & Haugen, T. (2015). Translating, reliability testing, and validating a Norwegian questionnaire to assess adolescents' intentions to be physically active after high school graduation. *Sage Open*, *5*(2), 2158244015580374.

Esterling, K. M., Brady, D., & Schwitzgebel, E. (2023). *The Necessity of Construct and External Validity for Generalized Causal Claims* (No. 18). The Institute for Replication (I4R).

Fink, A., & Kosecoff, J. (1985). How to conduct surveys Newbury Park.

Fleiss, J. L., Levin, B., & Paik, M. C. (1981). Statistical methods for rates and proportions. John Wiley & Sons. *New York*, *870*.

Forza, C. (2002). Survey research in operations management: a process-based perspective. *International journal of operations & production management*, *22*(2), 152-194.

Hahs-Vaughn, D. L., & Lomax, R. (2020). *An introduction to statistical concepts*. Routledge.

Institute of Medicine of the National Academies. (2015) . *Psychological Testing in the Service of Disability Determination*. National Academies Press (US).

Kember, D., & Leung, D. Y. P. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education*, *33*(4), 341–353.

Kerlinger, F. N. (1966). Foundations of behavioral research.

Keyton, J., King, T., Mabachi, N. M., Manning, J., Leonard, L. L., & Schill, D. (2004). Content analysis procedure book. *Lawrence, KS: University of Kansas*.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy*, *65*(23), 2276-2284.

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Knodt, A. R., Elliott, M. L., Whitman, E. T., Winn, A., Addae, A., Ireland, D., ... & Hariri, A. R. (2023). Test–retest reliability and predictive utility of a macroscale principal functional connectivity gradient. *Human brain mapping*, *44*(18), 6399-6417.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2225-2230

Kuder, G.F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151-160.
Liang, Y., Lau, P. W., Huang, W. Y., Maddison, R., & Baranowski, T. (2014). Validity and reliability of questionnaires measuring physical activity self-efficacy, enjoyment, social support among Hong Kong Chinese children. *Preventive medicine reports*, *1*, 48-52.
Lillis, A. (2006). Reliability and validity in field study research. *Methodological issues in accounting research: Theories and methods*, 461-475.
Liu, Y., Suh, K., Maini, P. K., Cohen, D. J., & Baker, R. E. (2024). Parameter identifiability and model selection for partial differential equation models of cell invasion. *Journal of the Royal Society Interface*, *21*(212), 20230607.
Lynn M. R. (1986). Determination and quantification of content validity. *Nursing research*, *35*(6), 382–385.
Mackison, D., Wrieden, W. L., & Anderson, A. S. (2010). Validity and reliability testing of a short questionnaire developed to assess consumers' use, understanding and perception of food labels. *European journal of clinical nutrition*, *64*(2), 210-217.
Miller, M. J. (2012). res 600: Graduate research methods: Reliability and validity. *Western International University.*
Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University. Economic Series*, *17*(4), 59-82.
Motl, R. W., Dishman, R. K., Trost, S. G., Saunders, R. P., Dowda, M., Felton, G., ... & Pate, R. R. (2000). Factorial validity and invariance of questionnaires measuring social-cognitive determinants of physical activity among adolescent girls. *Preventive medicine*, *31*(5), 584-594.
Nunnally, J.C. & Bernstein, I.H., Psychometric Theory, 3rd Ed. (New York: McGraw Hill, 1994).
Oluwatayo, J. A. (2012). Validity and reliability issues in educational research. Journal of Educational and Social Research, 2(2), 391-400.
Onwuegbuzie, A. J. (2000). Expanding the framework of internal and external validity in quantitative research.
Parsian, N., & Dunning, P. (2009). Developing and validating a questionnaire to measure spirituality: A psychometric process.
Patino, C. M., & Ferreira, J. C. (2018). Internal and external validity: can you apply research study results to your patients?. *Jornal brasileiro de pneumologia*, *44*(03), 183-183.
Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health*, *29*(5), 489-497.
Price, P., Jhangiani, R., Chiang, I, Leighton, D, Cuttler, C. (2017). Research methods in psychology. PB Pressbooks.
Sangoseni, O., Hellman, M., & Hill, C. (2013). Development and validation of a questionnaire to assess the effect of online learning on behaviors, attitudes, and clinical practices of physical therapists in the United States regarding evidenced-based clinical practice. *Internet Journal of Allied Health Sciences and Practice*, *11*(2), 7.
Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*. Pearson education.
Schaaf, J. V., Weidinger, L., Molleman, L., & van den Bos, W. (2024). Test-retest reliability of reinforcement learning parameters. *Behavior research methods*, *56*(5), 4582–4599.
Schimmack, U. (2010). What multi-method data tell us about construct validity. *European Journal of Personality: Published for the European Association of Personality Psychology*, *24*(3), 241-257.
Setia M. S. (2017). Methodology Series Module 9: Designing Questionnaires and Clinical Record Forms - Part II. *Indian journal of dermatology*, *62*(3), 258–261.
Shekhar Singh, A. (2014). Conducting case study research in non-profit organisations. Qualitative market research: an international journal, 17(1), 77-84.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.
Singh, A. S., Vik, F. N., Chinapaw, M. J., Uijtdewilligen, L., Verloigne, M., Fernández-Alvira, J. M., ... & Brug, J. (2011). Test-retest reliability and construct validity of the ENERGY-child questionnaire on energy balance-related behaviours and their potential determinants: the ENERGY-project. *International Journal of Behavioral Nutrition and Physical Activity*, *8*, 1-12.
Sugiyono. (2014). Statistika untuk penelitian. Bandung: Alfabeta.
Tabachnick, B. G. (2007). Using multivariate statistics (Vol. 5).
Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). sage.
Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53.
Thanasegaran, G. (2009). Reliability and Validity Issues in Research. *Integration & Dissemination*, *4*.
Thompson, B. (2003). Understanding Reliability. *Score reliability: Contemporary thinking on reliability issues*, 1.
Wahyudi, R. (2017). Uji validitas dan reliabilitas dengan pendekatan konsistensi internal kuesioner pembukaan Program Studi Statistika Fmipa Universitas Bengkulu. *FMIPA Universitas Bengkulu*, *1*, 105-112.
Wells, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. *Testing & Evaluation Services University of Wisconsin*.
Wong, K. L., Ong, S. F., & Kuek, T. Y. (2012). Constructing a Survey Questionnaire to Collect Data on Service Quality of Business Academics. *European Journal of Social Sciences*, *29*(2), 209-221.
Wu, D., Bednarczyk, C., RamonFigueroa, A., Zhu, H., Geer, M., Rosedale, R., & Robbins, S. M. (2025). Inter-rater reliability of Mechanical Diagnosis and Therapy (MDT) in evaluating and classifying chronic pelvic pain syndrome. *Journal of Manual & Manipulative Therapy*, 1-8.
Yarnold, P. R. (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: emergency severity index (version 3) and Canadian triage acuity scale. *Optimal Data Analysis*, *3*(4), 50-54.
Yusup, F. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1).
Zafrullah, Z., Atika Miftah Ramadhani, Rizki Tika Ayuni, Nuansa Trimaya Fadhilla, & Rina Safitri. (2024). The Using Confirmatory Factor Analysis as Construct Validity in Education Research: A Analysis with Biblioshiny. *DIROSAT: Journal of Education, Social Sciences & Humanities*, *2*(3), 206–220.