Psychology and Behavioral Sciences
2024, Vol. 13, No. 6, pp. 137-141
https://doi.org/10.11648/j.pbs.20241306.11

Methodology Article

Importance of Reliability and Validity in Research

Richard Karnia*

Department of Psychology, Elgin Community College, Elgin, United States

Abstract

The goal developing a new research tool is to ensure that the measurement tool has a high level of external validity to be generalizable and have a broader reach and also is highly reliable and able to consistently gather the same result. Researchers need to determine the validity and reliability of each assessment to ensure that they are not misleading their readers and the data can be trusted based on statistical evidence to support their conclusions. Reliability is the ability of consistency of the results over multiple tests. This process can be calculated by determining various measurements such as test-retest reliability, parallel-form reliability, split-half reliability by calculating a correlation coefficient or a t-test. Validity is the extent in which a test will measure what is said to test, which can be established by looking and measuring face validity, content validity, criterion-related validity, and construct validity. Validity can be established by using various experts to determine if a test is clear and relevant using a tool such as content validity index. If statistically reliability and validity is established, the research will increase the impact on the research and generalizability can be established.

Keywords

Reliability, Validity, Correlation Coefficient, Test-Retest Reliability, Content Validity Construct Validity, Content Validity Index

1. Introduction

Psychological assessment tools are created to use as a way for researchers to gather and integrate related data. An effective assessment tool can be used for the purpose of making a psychological evaluation or evaluating the validity and reliability of hypothesis being tested through a means of a devices or procedures designed to obtain a sample of behavior [3] The choice and design of a psychological tests and other tools of assessment will differ in terms of content, format, administration procedures, scoring and interpretation procedures, and technical quality [3]. With the implementation of a new or existing assessment tool it should continually be evaluated to ensure that a high level of validity and reliability are being maintained to ensure the effectiveness of the chosen evaluation tool.

2. Use of Reliability and Validity

The goal of any tool being used in psychological testing is to make sure that will effectively measure the values of all variables included in the design. The research design should include an effective way as part of the study to measure the reliability, its accuracy, and validity, the level of measurement it represents [2]. The proper evaluation of a research design and assessment tool's reliability and validity are important steps as it helps ensure that the research will have a greater level of generalizability, and can help limit and reduce confounding variables impacts on the measurement tool. If accounted for, confounding variables do not always result in a threat to internal validity when they are identified, and shown to have little to or no effect on the dependent or criterion viable and can be taken into account in the analysis [2].

*Corresponding author: Karnia.richard@elgin.edu (Richard Karnia)

Received: 26 September 2024; **Accepted:** 23 October 2024; **Published:** 12 November 2024

When reliability and validity are established, generalization can be established by applying the findings to a larger population and across a variety of research settings [2]. The following discussion will be evaluated techniques that will ensure ways to effectively measure quantitatively and qualitatively the reliability and validity of research question or assessment tool.

2.1. Reliability and Reliability Coefficient

Reliability is the process that concerns the ability a measurement of variables to be able to produce the same results and measurements under repeated conditions [2]. Reliability will measure the consistency, precision, repeatability, and trustworthiness of research and indicates the extent to which it is without bias [5]. Reliability is used in qualitative research and is the degree to which an assessment tool is free from errors, produces consistent results, and is a necessary component of validity [5]. In quantitative research reliability is the consistency, stability, and repeatability of results that are obtained in identical situations, but different circumstances such as across different researchers in different projects [5]. The reliability coefficient is a statist that will quantify reliability and can be uses to measure test-retest reliability, alternate form reliability, split half reliability, and inter-score reliability [3]. High reliability is based on the correlation coefficient between two variables with the data falling between a 0, which indicates no reliability, and a 1 which will indicate perfect reliability [5]. For high stakes measurements the reliability should be greater than 0.9, but general rule that reliability greater than 0.8 is considered high [5]. The ability to use statistics to accurately measure the reliability of test will allow research to have more confidence in using new testing methods. Reliability scores will show the stability of the measures being tested at different times over the same items, and the more reliability found in the results, the greater accuracy of the data, which will lead to an increase chance of making the correct decision in research [5].

2.2. Test Developers and Measurement Tools of Reliability

Test developers can ensure the reliability of test by examining multiple measures of reliability. Test-retest reliability is when a construct being measured is consistent across time and the scores being obtained remained consistent [12]. Measuring a construct using test-retest reliability requires that the measure was used on a group of people will be used again on the same group of people at a later time and the data will be graphed to determine the relation of the correlation coefficient showing good reliability will have a score of +.80 [12] It is important determine a set period of time between the two measurements ranging from a few weeks to a few months and the retest with a low correlation can indicate that too much time has occurred, maturation has taken place, or there are errors in the measurement [5]. An example of test-retest could be a personality measurement that when given several weeks or months later will result in the same responses/score which would indicate a high level of reliability [1] Parallel-forms reliability is a measure of the reliability that is obtained in research when two different assessments tools are administered to the same group of individuals and the scores from the two versions are correlated to evaluate any relationship between the two versions [5]. Items that are assess on a parallel-form are supposedly equivalent to the items that are found in the original form, assess the same knowledge and skills but will use different questions or problems to eliminate the possibility that the person is just recalling an answer from a previous assessment [2]. Split-half reliability is designed to measure the internal consistency by checking one half of the results scaled against the other half of an assessment [5]. A split-half reliability can be done to compare the first half and second half or even and odd numbers, and if similar results are found this would indicate the test has internal validity [5]. The internal reliability is the consistency of people's responses across responses on a multiple item measure or multiple-choice assessment in which all measures are supposed to reflect the same underlying construct so peoples scores on the two halves should be correlated with each other [12]. Internal reliability will result in getting a consistent measurement/result from different parts of a test that is all measuring the same thing [1]. An example would be a questionnaire that focuses on self-esteem. The assessment will be split into two halves, if the results of each half are the same there is a high level of reliability, but if the results varied in the two halves it would indicate a low internal consistency and low reliability of the two halves [1]. When measuring the data between various factors of reliability a t-test should be used to identify if there are any statistical differences between the two scores [7].

2.3. Test-Retest Reliability Challenges and Constraints

Test-Retest Reliability is important to establish the utility of the assessment tool and being able to predict generalizable outcomes [14]. A prerequisite of test-retest reliability is to establish parameter identifiability to ensure stability over time [14]. Parameter identifiability is the process of identifying the extent that all parameters of a model can be estimated from the available data [8]. Parameter identifiability has an impact on the reliability of any model and if they are not pinned down can lead to a model prediction that may lead to different results under a different sent of conditions which would reduce the generalizability of the study [8]. If the parameters can not be identified it will lead to noise in the data collection which could then harm the test-retest reliability [14]. Several other factors that can influence the test-retest reliability is the if the participants have stable behaviors, measuring if people are different across a period of

time in cognition and moods, and if traits were accurately measured in the original assessment [14].

2.4. Validity and Validation

Validity is a measure the extent in which the research or measurement measures what it was intended to measure [2]. When designing an experiment, the goal is to establish a good test-retest reliability, internal consistency which will then allow researchers to be confident that the scores represent what they are supposed to, which shows validity of the research [12]. Validity is a degree to which the instruments measures what it was designed to and the extent in which the results are truthful [5]. To help ensure validity of the results the instruments or test used has to encompass the full experimental concept and meet all the requirements of the scientific method [5]. The analysis of the data being gathered by the test and evaluating the validity is validation [3]. Qualitative research validity is based on the extent of how the scientific method has been followed while conducting research to generate results that will demonstrate a matter of trustworthiness, utility, and dependability that is based on accuracy of the instruments scores and interpretations [5]. In quantitative research the validity is based on the extent a measuring device or test accurately measures what it is intended to measure [5]. Validity in research is composed two parts, internal validity and external validity. Internal validity is the extent in which a study is legitimate based on the way the sample group was selected, data was recorded and analyzed [5]. External validity will show if the results are transferable or that the results of the given study can be transferable to other groups of interests [5]. If there is no external validity a researcher will risk not understanding the cause to have an effect and external validity is needed to make a generalizable claim about cause and effect [4]. A lack of internal validity will imply that the results have deviated from the truth and one cannot draw any conclusions and external validity is irrelevant [11]. To increase validity, one must carefully plan all research for quality control by impending strategies that include adequate recruitment strategies, data collection, data analysis, sample size and include criteria that resembles real life situations [11].

For a test to be considered accurate it must be reliable and have a high level of internal validity. The measure must have been developed with sound measure, explicit methods and procedures by which tasks should be administered are determined and clearly spelled out which is also known as standardization [6]. Several key elements that produced a standardized testing environment can include a quiet, relatively distraction-free environment, a reading of scripted instructions, and all required tools or stimuli [6]. Highly accurate tests that standardized tests provide a set of normative data to which an individual's performance can be compared [6]. The norms should be based upon representative samples of individuals from the intended test population, as each person should have an equal chance of being in the standardization sample and when a test is applied to individuals for whom the test was not intended or not included as part of the norm group, inaccurate scores and subsequent misinterpretations may result [10].

2.5. Test Developers and Measurement Tools of Validity

A test developer will use various measures to ensure the validity of the assessment that includes face, content, criterion-related, and construct validity. When selecting a measurement tool, the validity should have been tested to measure content and construct variables to make sure all data being gathered is relevant and accurate [16]. Validity is explained as the relevancy between the evidence and theories that allows for a score interpretation aligned with the purpose of the test [13]. Face validity is the subjective assessment of a questionnaire that is often done by a subject matter expert after designing an instrument and evaluate if it appears to be appropriate and has relevant items on the "face of it" [15] Face validity is at its best a very week kind of evidence that a measurement method is measuring what it is supposed to [12]. A content validity is a form of subjective assessment when researcher will assess whether a questionnaire adequately measures the concepts they are supposed to [15] Content validity will determine if the questions on the and the scores on the assessment represent all possible questions that could be asked about the content or skill and will be able to assess current performance rather than predicting future performance [5]. A criterion related validity is when a measure is designed to look at specific outcomes that match with an existing standard [15] This type of measurement is used to predict future performance and is often demonstrated with a relationship between scale scores and a specific measurable criterion as the future performance will be based on scores currently obtained by the measure and how they correlate to the scores obtained with the performance [5]. Construct validity is when a researcher will try to examine the items within a questionnaire in respect to the underlying hypothesis [15] These variables that are being measured are not directly observable and has been developed to explain a behavior [2]. When using a construct validity, one must demonstrate that those who scored high or low on a measure behave as predicted by the theory [2].

2.6. Measuring Content Validity

Content validity can be determined by gathering data from each assessment tool being used and determining the content of the test and the construct being measured [13]. The various parts of the test that contribute to the content include the wording, format, display of items, and a test is only valid when the items relevantly measure the construct [13]. When designing a test, the following steps should be taken to ensure content validity, the definition and information measured by the construct, the number of items for each aspect or dimension, and item test arrangement [13]. When evaluating

the content validity an assessment tool the researcher will want to ensure the tools representativeness, clarity, factor structure, and comprehensiveness [13]. Once a tool has been created an item evaluation should be completed by five to ten experts to determine which items should be kept or removed from the testing tool [13].

An analysis can be done using a content validity index Following Completion of the evaluation of the exam by each member of a subject matter teams or department chair independently, the data will be organized using the content validity index which show if each item should be revised, removed, or is valid. The content validity index is used as a way to organize data and quantifiably summarize item relevancy score from a panel of experts [9]. The content validity assessment is calculated by counting the number individuals who give a test question a 3 or 4 score (on a 1-4 scale) for relevance or clarity on each assessment item and then divide it by the number individuals who evaluated the assessment [17]. If 10 members of a team are evaluating a test, and of seven individuals give a score of 3 or 4 to an item, the Content Validity Index would be: 7/10 = .700 (I-CVI = .700). After calculating the CVI for each assessment item, if the I-CVI is below .7 (I-CVI < .70) the test or assessment question should be removed, test or assessment question between .70 and .90 (.70 ≤ I-CVI ≤ .90) should be revised, and items with a Content Validity Index above .90 (I-CVI > .90) should remain. When measuring the content validity of an assessment it is difficult to have a unanimous commences among all members, and data being gathered can be affected if there are too few or too many items, the suggest number of evaluators should be between five to ten [13].

3. Conclusion

Without ensuring the proper validity and reliability of the assessment tool or research design, the results of any data gathered will be questionable, lack of external validity, and will reduce the generalizability of your research. Establishing reliability in research will help ensure the consistency, repeatability, and trustworthiness of a research design and create confidence in the use of the assessment tool. Test-retest reliability needs should try to ensure that the models parameters are parameter identifiability. The confirmation of validity of the assessment tool will allow the researcher to have a degree of confidence the scores represent what they are supposed to and establish a level truthfulness through criterion and construct validity. The generalizability of the research tool to a larger population will increase the impact of one's research as a result of the establishment of external validity. Psychological tests and research design are only useful to the extent that they really measure the characteristics they claim to measure which can only be established by using the discussed methods and ensuring a high level of reliability and validity.

Abbreviations

CVI Content Validity Index

Author Contributions

Richard Karnia is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Ahmed, I. & Ishtiaq, S. (2021, October). Reliability and validity importance in medical research. *Journal of the Pakistan Medical Association.*

[2] Bordens, K. & Abbott, B. (2014). *Research design and methods a process approach*. McGraw- Hill Education.

[3] Cohen, R., Schneider, W., & Tobin, R. (2022). *Psychological testing and assessment an introduction to tests and measurements (10th ed.).* McGraw Hill.

[4] Esterling, K., Brady, D., Schqitzgebel, E. (2023, February). The necessity of construct and external validity for generalized casual claims. *Institute for Replication.*

[5] Haradhan, M. (2017, October 1). Two criteria for good measurements in research: validity and reliability. *Annals of Spiru Haret University.*

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

[7] Knodt, A., Elliot, M., Whitman, E., Winn, A., Addae, A., Ireland, D., Poulton, R., Ramrakha, S., Caspi, A., Moffitt, T., Hariri, A., (2023, September). Test-retest reliability and predictive utility of a macroscale principle function connectivity gradient. *Wiley Periodicals LLC.* https://doi.org/10.1002/hbm.26517

[8] Liu, Y., Suh, K., Maini, P., Cohen, D., Baker, R. (2024, January 31). Parameter identifiability and model selection for partial differential equation models of cell invasion. J. R. SOC. Interface 21. https://doi.org/10.1098.rsif.2023.0607

[9] McCoach, D. B., Gable, R. K., & Madura, J. P. (2013*). Instrument development in the affective domain (3rd ed.).* Springer.

[10] National Library of Medicine (2015, June 29). Committee on Psychological Testing, Including Validity Testing, for Social Security Administration Disability Determinations; Board on the Health of Select Populations; Institute of Medicine. Washington (DC): National Academies Press (US); https://www.ncbi.nlm.nih.gov/books/NBK305230/Na

[11] Patino, C. & Ferreira, J., (2018) Internal and external validity: can you apply research study results to your patients? *Sociedade Brasileira de Pneumologia e Tisiolgia*.

[12] Price, P., Jhangiani, R., Chiang, I, Leighton, D, Cuttler, C. (2017). *Research methods in psychology*. PB Pressbooks.

[13] Roebianto, A., Savitri, S., Aulia, I., Suciyana, Mubarokah, L., (2023, March). Content validity: definition and procedure of content validation in psychological research. *TPM*. 30. 1.

[14] Schaaf, J., Weidinger, L., Molleman, L., Bos, W., (2023, 8 September). Test-retest reliability of reinforcement learning parameters. *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02203-4

[15] Setia, M. (2017 May-June). Methodology series module 9: designing questioners and clinical records forms - part II. *Indian Journal of Dermatology.*

[16] Zafrullah, Z., Ramadhani, A., Ayuni, T., Fadhilla, T., Safitri, R. (2024). The using confirmatory factor analysis as construct avalidty in education research: a analysis with biblioshiny. *Journal of Education, Social Science and Humanities* 2. 3 https://doi.org/10.58355/dirosat.v2i3.70

[17] Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, 4(2), 165-178. https://doi.org/10.15171/jcs.2015.017