ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2617-9415 (Online)
CODEN: ESSDAX

Education, Sustainability & Society (ESS)

DOI: http://doi.org/10.26480/ess.01.2023.36.40

RESEARCH ARTICLE

A POINTER TO VALIDITY, TYPES, AND POSSIBLE THREATS TO INSTRUMENT DESIGN AND IMPLEMENTATION IN RESEARCH PRACTICES AND CLASSROOM ASSESSMENT

Eyong, Emmanuel Ikpi

Department of Educational Foundations and Childhood Education, Faculty of Education Cross River University of Technology (CRUTECH), Calabar-Nigeria.
**Corresponding Author Email: piusofem396@gmail.com*

ARTICLE DETAILS

Article History:

Received 06 January 2023
Revised 08 February 2023
Accepted 10 March 2023
Available online 14 March 2023

ABSTRACT

The study examines a pointer to validity, types, and possible threats to instrument design and implementation in research practices and classroom assessment. The thrust of the study was based on the fact that in instrument design and implementation, validity is a sure tool that influences the quality of an instrument. This is because an instrument that is not validated is bound to generate baseless results. This call for instrument designers to have a better understanding of the types and possible threat that could limit the generalizability of a research instrument. In this light of this, the present study sought to research a glimpse into validity, types, and possible threats to instrument design and implementation in research practices. The study pinpoints the need for validity, different types, and possible threats to could hinder the legality, authenticity, and legitimacy of any research instrument meant for data collection. It is not rhetoric that a research topic may be well articulated, with a well-stated objective, research questions, hypotheses, and well-reviewed literature with a well-stated design and methodology. However, if the instrument is not well validated by experts in the field of study, there are bound to be faulty findings that will result in baseless (senseless) generalization. This underscores the need for scholars to have a better pointer to the possible internal and external threats to the validity of a research instrument, hence, the thrust of the present study.

KEYWORDS

Pointer Validity, Threats, Instrument Design, Research Practices, Classroom Assessment

1. INTRODUCTION

Research practices cannot be appropriately executed if the instrument fails the test of validity. To this end, to establish the usefulness, meaningfulness, and appropriateness of any instrument design to be implemented, it must pass the phase of validation. This is because, without validity, the aim of research practices cannot be adequately accomplished as such generalization will be baseless due to the faulty nature of the data collected from the instrument. The concept of validity was propounded by who rightly said that a test is valid if it measures what it purports or intended to measure at its inception (Joshua, 1998; Kelly, 1927). For instance, a test designed to measure mathematics ability should measure Mathematics ability alone without interference in other aspects that the test did not intend to measure at its inception. If such a test goes to measure students' Physics or Chemistry ability then it can be concluded that the test is not valid (invalid). Mathison rightly defined validity as the extent to which a test measures what it purports (claims) to measure (Mathison, 2005). It refers to the credibility or believability of the research instrument. To this end, validity in the language of students is the distinction between a "fair" examination and an "unfair" examination.

This implies that a fair examination measures what it purports to measure, namely, the student's knowledge and understanding of the subject matter; and an unfair examination are one for which the student's score substantially reflects something other than knowledge and understanding, for example, the student's ability to spot and deal with trick questions, or to adhere to some particular theoretical or ideological party line favored by the instructor (Crombach, 1970). The three questions always raised in ascertaining the validity of an instrument are; the form of the test, the purpose of the test, and the population for whom it is intended. Therefore, we cannot ask the general question: Is this a valid test? How valid is this test for the decision that it needs to make? Are the findings genuine? There are several types of validity namely face validity, content validity, construct validity, and criterion-related validity (predictive and concurrent validity); internal, external, multi-trait-multi methods.

2. TYPES OF VALIDITY OF RESEARCH INSTRUMENTS

There are several types of instrument validity which include face validity, content validity, construct validity, Multi-trait-Multi-method (MTMM), criterion validity, and internal and external validity of a research instrument. These different types and painstakingly explained below:

2.1 Face Validity of Research Instruments

Face validity is the least sophisticated measure of validity. It simply measures whether the test appears (at face value) to measure what it claims. Face refers to the identity of anything, as such in Measurement and Psychology, face validity is the degree to which an assessment or test subjectively appears to measure the variable, construct, trait, or feature that it intends to measure. Face validity refers to the superficial, physical, or outward appearance (mere facial outlook) to which a procedure,

Quick Response Code	Access this article online	
	Website: www.educationsustability.com	DOI: 10.26480/ess.01.2023.36.40

Cite The Article: Eyong, Emmanuel Ikpi (2023). A Pointer to Validity, Types, and Possible Threats to Instrument Design and Implementation in Research Practices and Classroom Assessment. *Education, Sustainability & Society, 6(1): 36-40.*

especially a psychological test or assessment, appears effective in terms of its stated aims. In other words, face validity can be derived when an assessment or test appears to do what it claims to do. Face validity determines the extent to which the results of the instrument are seen based on what they look like.

In this regard, we look at how valid a measure appears on the surface and make subjective judgments based on its superficial appearance. In research, it is never enough to rely on face judgments alone and more quantifiable methods of validity are necessary to draw acceptable conclusions. For example, if a mathematics teacher claims that his test measures the mathematical ability of his students. Since all of the students who took the test agreed that the test appears to measure mathematical ability, then it can be concluded that the test has face validity. It is important to know that face validity does not necessarily mean that a test is a valid measure of a construct, but rather, the test looks like it is a valid measure.

2.2 Content Validity of Research Instruments

Content validity occurs when the instrument provides adequate coverage of the subject being studied. This includes measuring the right things as well as having an adequate sample. Samples should be large enough and be taken for appropriate target groups. The perfect question gives a complete measure of all aspects of what is being investigated. A high content validity question covers more of what is sought. Content validity ensures that all of the target content is covered (preferably uniformly). Content validity deals with whether the content and composition of the instrument are appropriate, given what is being measured. For example, does the test content reflect the knowledge/skills required to demonstrate that one grasps the course content sufficiently? Content validity is whether or not the measure used in the research covers all of the content in the underlying construct (the thing you are trying to measure). This is also a subjective measure, but unlike face validity, we ask whether the content of a measure covers the full domain of the content. If a researcher wanted to measure mathematics anxiety, they would have to first decide what constitutes a relevant domain of content for that trait. Where content validity distinguishes itself (and becomes useful) through its use of experts in the field or individuals belonging to a target population. This study can be made more objective through the use of rigorous statistical tests. For example, you could have a content validity study that informs researchers how items used in a survey represent their content domain, how clear they are, and the extent to which they maintain the theoretical factor structure assessed by the factor analysis.

2.3 Construct Validity of Research Instruments

Construct validity was invented by as they refer to the extent to which a test captures a specific theoretical construct, characteristic, or trait (s) it intends to measure (Cornball and Meehi, 1955). Construct validity does not concern the simple, factual question of whether a test measures an attribute. Instead, it is about the complex question of whether the test score interpretations are consistent (Cronbach and Meehl, 1955). To determine the construct validity of a research instrument, it must be demonstrated that the phenomenon being measured exists. For instance, the construct validity of an intelligence test is dependent on a model or theory of intelligence. The more evidence a researcher can demonstrate for a test's construct validity the better. However, there is no single method of determining the construct validity of a test. Instead, different methods and approaches are combined to present the overall construct validity of a test. For example, factor analysis and correlational methods can be used. Two types of construct validity exist namely; convergent and discriminant validity.

To a construct represents a collection of behaviors that are associated in a meaningful way to create an image or an idea invented for a research purpose (Monday, 2005). Construct validity is the degree to which your research measures the construct (as compared to things outside the construct). Depression is a construct that represents a personality trait that manifests itself in behaviors such as oversleeping, loss of appetite, difficulty concentrating, etc. Construct validity is the degree to which inferences can be made from operationalizations (connecting concepts to observations) in your study to the constructs on which those operationalizations are based. To establish construct validity you must first provide evidence that your data supports the theoretical structure. You must also show that you control the operationalization of the construct, in other words, show that your theory has some correspondence with reality.

2.4 Multi-Trait-Multi-Method (MTMM) Research Instruments

Campbell and Fiske define MTMM as a form of validity that demonstrates construct validity by using multiple methods. For instance, surveys, observations, tests, etc measure the same set of 'traits' and show correlations in a matrix, where blocks and diagonals have special meanings. Multi-trait-Multi-method (MTMM) designs refer to a construct validation approach proposed by Campbell and Fiske in 1959. To apply MTMM designs, researchers assess multiple traits (i.e., psychological constructs) for a group of individuals using multiple methods that are maximally different. Correlation coefficients among the multiple constructs so produced are then compared to evaluate convergent and discriminant validity. To ensure validity, correlations between the same traits assessed with different methods must be sufficiently large and larger than those between different traits assessed with either the same or different methods. Further, the same pattern of correlations should exist between traits within each method (Campbell and Fiske, 1959).

Although the Multi-trait-Multi-method (MTMM) approach as a standard technique for construct validation (Campbell, 1960; Campbell and Fiske, 1959), seeks to establish higher correlations across diverse measures of the same trait (convergent evidence) and lower correlations among similar measures of different traits (discriminant evidence) to show that a construct is distinct from other constructs and that it is not uniquely tied to a particular measurement method.

In Multi-trait-Multi-method (MTMM), convergent validity occurs where measures of constructs are expected to correlate and correlate perfectly. That is the degree to which an operation is similar to other operations it should theoretically be similar to. This is similar to concurrent validity (which looks for correlation with other tests). Put in a more specific way, In convergent validity, the focus is to examine the degree to which the operationalization is similar to (converges on) other operationalizations that it theoretically should be similar to. For instance, to show the convergent validity of a programme that says a sandwich programme at the University of Calabar, we might gather evidence that shows that the programme is similar to other Sandwich programs. Put differently, to show the convergent validity of a test of Statistics skills, we might correlate the scores on our test with scores on other tests that purport to measure Mathematics skills, where high correlations would be evidence of convergent validity.

On the other hand, discriminant validity occurs where constructs that are expected not to relate do not, such that it is possible to discriminate between these constructs. If a scale adequately differentiates itself or does not differentiate between groups that should differ or not differ based on theoretical reasons or previous research. In discriminant validity, the instrument examines the degree to which the operationalization is not similar to (diverges from) other operationalizations that it theoretically should be similar to. For instance, to show the discriminant validity of a Sandwich programme at the University of Calabar, we might gather evidence that shows that the programme is *not* similar to other programmes that do not label themselves as a Sandwich programme. Put differently, to show the discriminant validity of a test in Statistics skills, may require one to correlate the scores on our test with scores on tests of communication skills, where *low* correlations would be evidence of discriminant validity.

2.5 Criterion-Related Validity of Research Instruments

The criterion-related validity of a test is used to compare a test to some external factors known as criteria. The criterion can be another test or even some type of outcome. Frequently the criterion is another test measuring close to the same thing as the test being evaluated is purported to measure. Criterion-related validity is further classified into either predictive validity or concurrent validity. Criterion-related validity (also called instrumental validity) is a measure of the quality of your measurement methods. The accuracy of a measure is demonstrated by comparing it with a measure that is already known to be valid. In other words, if your measure has a high correlation with other measures that are known to be valid because of previous research. For this to work you must know that the criterion has been measured well. And be aware that appropriate criteria do not always exist. What you are doing is checking the performance of your operationalization against criteria. The criteria you use as a standard of judgment account for the different approaches you would use:

2.6 Predictive Validity of Research Instruments

In predictive validity**,** we assess the operationalization ability to predict something it should theoretically be able to predict. *This* measures the extent to which a future level of a variable can be predicted from a current measurement. This includes correlation with measurements made with different instruments. This is the degree to which a test accurately predicts a criterion that will occur in the future. For example, a prediction may be

made based on an intelligent student who performed in an intelligence test that high before schooling one can comfortably predict that such a child will do well academically in school. If the prediction is born out then the test has predictive validity.

2.7 Concurrent Validity of Research Instruments

Concurrent validity is the degree to which the scores on a test are related to the scores on another, already established, a test administered at the same time, or to some other valid criterion available at the same time. In concurrent validity**,** we assess the operationalization ability to distinguish between groups that it should theoretically be able to distinguish between. These measures the relationship between measures made with existing tests. The existing tests are thus the criterion. A measure of creativity should correlate with existing measures of creativity. For example, if a new simple test is to be used in place of an old and cumbersome one. Concurrent validity is the degree to which a test corresponds to an external criterion that is known concurrently (i.e. occurring at the same time). If the new test is validated by comparison with a currently existing criterion, we have concurrent validity.

2.8 Internal Validity of Research Instruments

Research is generally conducted to determine cause-and-effect relationships. That is, a change in the independent variable caused the observed changes in the dependent variable. If a study shows a high degree of internal validity then we can conclude that there is strong evidence of causality. On the other hand, if a study has low internal validity then can conclude that there is little or no evidence of causality of the instrument. According to Kothari (2004), internal validity is the extent to which observed differences in the dependent variable are directly related to the independent variable. If a relationship is observed that is not related to extraneous variables such as differences in subjects, location, or other related factors, the research probably has strong internal validity. Internal validity occurs when it can be concluded that there is a causal relationship between the variables being studied. A danger is that changes might be caused by other factors. It is related to the design of the experiment, such as in the use of random assignment of treatments. Internal validity refers to the extent to which the independent variable can accurately be stated to produce the observed effect. If the effect of the dependent variable is only due to the independent variable(s) then internal validity is achieved. This is the degree to which a result can be manipulated. Put another way, internal validity is how you can tell that your research "works" in a research setting. Within a given study, does the variable you change affect the variable you're studying?

2.9 The External Validity of Research Instruments

External validity occurs when the causal relationship discovered can be generalized to other people, times, and contexts. Correct sampling will allow generalisation and hence give external validity. External validity refers to the extent to which the results of a study can be generalized beyond the sample (Monday, 2005). This is to say that you can apply your findings to other people and settings. External validity consists of a determination of whether the results of the experiment can be generalized to an entire population from which the samples were drawn in the study. External validity tells us the degree to which the results of an empirical investigation can be generalized to and across individuals, settings, and times.

3. Possible Factors That Affect Internal and External Validity of An Instrument

In establishing the validity of an instrument several factors can pose threat or affect the validity and reliability of a research instrument and they are internal and external factors. Internal validity is even more basic since it refers to whether it can be concluded that the independent variable produced the differences observed. The matter of external validity is secondary to and dependent upon the demonstration of adequate attention to the threats to internal validity (Campbell and Stanley, 1963).

3.1 Internal Factors of Research Instruments

The following can pose threats to the internal validity of an instrument:

i. Subject characteristics. The subject characteristics threat occurs when the selection of subjects results in differences between groups that are related to the different variables being studied. This difference can cause a selection bias. The selection bias means that something in the makeup of the subjects favours one group or another group.

ii. Location. The location threat means that something about the setting or settings of the study affects the outcome, either positively or negatively. Many factors could result in a location threat, and in a research study, these factors may influence the results. These differences in location could include differences in technology between two groups, differences in teacher or staff morale, and many other factors. To minimize these threats of location, the researchers should try to implement the program in a way that ensures the least possible differences in the locations used in the study.

iii. History: The specific events which occur between the first and second measurement of behavior at different points in time could result in differences reflecting the impact of the independent variable or extraneous and unwanted effects occurring as a result of cultural change (war, famine) over which the experimenter has no control. History is a threat to conclusions drawn from longitudinal studies. The greater the period elapses between measurements, the more the risk of a history effect.

iv. Maturation: The processes within subjects that act as a function of the passage of time. That is, if the test lasts a few years, most participants may improve their performance regardless of treatment. This may produce changes across time (nervous system growth or becoming fatigued) which can produce behavioral changes unrelated to experience or the impact of an experimental variable. Thus, the vital and continuing issue of the nature-nurture controversy is evident.

v. Repeatedly testing participants using the same measures influences outcomes. If you give someone the same test three times, isn't it likely that they will do better as they learn the test or become used to the testing process so that they answer differently?

vi. Testing effects: This consists of either reactivity as a result of testing or practice/learning from exposure to repeated testing. Longitudinal studies which require participants to take certain tests on several occasions are subject to this threat to internal validity.

vii. Selection: It refers to selecting participants from the population for the various groups in the study. In the groups' equivalents at the beginning of the study? If subjects were selected by random sampling and random assignment, all had an equal chance of being in treatment or comparison groups, and the groups are equivalent. Were subjects self-selected into experimental and comparison groups? This could affect the dependent variable. Selection is not a threat to the one-group design but it is a threat to the two-group design.

viii. Statistical regression: Statistical regression occurs where repeated measures are used and are particularly evident when participants are selected for study because they are extreme on the classification variable of interest, e.g. intelligence. Test-retest scores tend to systematically drift to the mean rather than remain stable or become more extreme. Regression effects may obscure treatment effects or developmental changes.

ix. Instrumentation problems: This is more of a concern in longitudinal studies where over significant periods researchers leave the study and testing instruments become invalid because of cultural change (tests are typically revised/re-normed every 10 years). Changes in experimenters may introduce different observers or techniques which could alter the continuity of measurement.

x. Instrumentation (instrument decay): This threat refers to how instruments are used in the study, which may cause a threat to the internal validity of the study. There are several ways in which an instrumentation effect may occur. One way is through what is referred to as instrument decay. The procedure for administering the instrument changes over time. For example, if a person is collecting data by making observations, they may start looking for different things over some time. This change can cause the results of the observations to change significantly. Another way in which an instrument may decay is through the fatigue of the person administering the instrument. As the researcher administers the instrument for data collection, at some point in data administration, he/she may get tired (fatigued), and then miss certain behaviors that are important for the study.

xi. The attitude of the subjects toward the instrument: This can be caused by something such as test fatigue of students taking the instrument. The students may feel that they are being tested too

often. This feeling could cause them to get tired of having to take another test and not do their best. This change in attitude could cause the results of the study to be biased.

xii. Experimental Mortality or attrition of subjects: This is a major threat to a lengthy longitudinal study since the sample remaining at the end of the study is unlikely to be comparable to the initial sample. Differential loss of participants across groups. It poses certain questions like. Did some of the respondents (participants) drop out? Did this affect the results? Did about the same number of participants make it through the entire study in both experimental and comparison groups? Mortality refers to the loss of subjects which can hinder the generalizability of the research and can also introduce bias.

xiii. Design contamination: Investigators must interview subjects after the experiment conclude to find out if design contamination occurred. it tries to answer the question: Did the comparison group know (or find out) about the experimental group? Did either group have a reason to want to make the research succeed or fail? John Henry effect: John Henry was a worker who outperformed a machine in an experimental setting because he was aware that his performance was compared with that of a machine.

xiv. Biases in sample selection: These are threats to both the cross-sectional and longitudinal approaches although, because of the cost and time commitment, they are devastating to the conclusions drawn from the longitudinal study.

3.2 Threats to The External Validity of An Instrument

There are many threats to the External validity of an instrument as pointed out by which are fully explained below (Gronlund, and Linn, 1985).

i. **Hawthorne Effect:** The Hawthorne Effect occurs due to the subject awareness of the existence of a phenomenon in the experiment. In our daily life, the inclination of people who are the subjects of an experimental study to change or improve their behavior is evaluated only because it is being studied and not because of changes in the experiment parameters or stimulus. The Hawthorne Effect is based on the fact that people will modify their behavior simply because they are being observed.

ii. Population validity: This determines how representative is the sample of the population. This is because the more representative, the more confident we can be in generalising from the sample to the population. It is also concerned with how widely the finding applies. Generalizing across populations occurs when a particular research finding works across many different kinds of people, even those not represented in the sample. Population validity relates to how well the experimental sample represents a population. The sampling methodology addresses this issue.

iii. Ecological validity: This has to do with the degree to which a result generalises across settings. Ecological validity relates to the degree of similarity between the experimental setting and the setting to which you want to generalize. The greater the similarity of key characteristics between settings, the more confident bet can be that the results will generalize to that other setting. In this context, "key characteristics" are factors that can influence the outcome variable.

iv. Generalizability requires that the methods, materials, and environment in the experiment approximate the relevant real-world setting to which you want to generalize. Threats to external validity are differences between experimental conditions and the real-world setting. Threats indicate that you might not be able to generalize the experimental results beyond the experiment. Research undertakings are performed in a particular context, at a particular time, and with specific people. As you move to different conditions, you lose the ability to generalize. The ability to generalize the results is never guaranteed. This issue is one that you need to think about. If another researcher conducted a similar study in a different setting, would that study obtain the same results? The following types of ecological validity include Interaction effects of testing, interaction effects of selection biases and experimental treatment, reactive effects of experimental arrangements, multiple-treatment interference, and experimenter effects.

v. Interaction effect of testing: Pre-testing interacts with the experimental treatment and can cause some effect, such that the results will not generalize to an untested population. Interaction effects of selection biases and the experimental treatment. In a physical performance experiment, the pre-test clues the subjects to respond in a certain way to the experimental treatment which would not be the case if there were no pre-test.

vi. Interaction effects of selection biases and the experimental treatment: The results of an experiment in which the teaching method is the experimental treatment, used with a class of low achievers, do not generalize to heterogeneous ability students.

vii. Multiple-treatment interference: When the same subjects receive two or more treatments as in a repeated measures design, there may be a carryover effect between treatments such the results cannot be generalized to single treatments. In an experiment with teaching methods, the same students are administered four different teaching methods. The effects of the second through fourth teaching methods cannot be separated from the possible delayed effects of the preceding method.

viii. Pre-post-test effects: When the pre-post-test is in some way related to the effect seen in the study, such that the cause-and-effect relationship disappears without these added tests.

ix. Sample features: When some feature of the particular sample was responsible for the effect (or partially responsible), leading to limited generalizability of the findings

x. Situational factors: Time of day, location, noise, researcher characteristics, and how many measures are used may affect the generalizability of findings

4. The Implication of the Study on Research Practices and Classroom Assessment

In the classroom context, the teachers' role is to employ set instruments with a high level of validity and reliability. This can only be achieved when the test employed is valid. To obtain useful results, the methods the teacher employs to collect data must be valid. This implies that the research must be measuring what it claims to measure. This ensures that your discussion of the data and the conclusions you draw are also valid. this assertion is supported by who collectively informed that the validity of a research instrument assesses the extent to which the instrument measures what it is designed to measure (Robson, 2011; Pallant 2011). It is the degree to which the results are truthful. So, it requires a research instrument (questionnaire or test) to correctly measure the concepts under the study. The role this will play in the assessment of the student is that it will clearly show the characteristics the test measures and how well the test measures that characteristic. Validity tells you if the characteristic being measured by a test is related to job qualifications and requirements expected in the test instrument (s). this will help the instructor to interpret the results as meaningful indicators of what we are trying to measure.

5. Summary of The Study

An instrument is a device used by a researcher in data collection. In instrument design and implementation, the concept of validity determines how useful, meaningful, and appropriate an instrument is for the purpose it was designed. Instrument construction and testing are a part of learning and let students display their understanding and ability of what has been taught to them in school. Test results show students' strengths and weaknesses in various subjects taught. This underscores why validity is necessary for the research and classroom assessment process. In summary, validity is a strong determining factor to be considered in instrument design. Thus, when an instrument for data collection is not properly validated it is doomed of producing faulty results and generalization. The study has outlined several types and possible factors both internal and external factors that can help researchers and instrument constructors to reduce threats in designing and implementing research instruments. It is worthy to summarize that every research undertaking that requires instrument validation should be adequately done to avoid false findings which can further aggravate false generalization in research skills and application.

References

Combach, L.J., 1970. Essentials of Psychological Testing (3rd Ed.) New York: Harper and Row.

Gronlund, N.E., and Linn, R.L., 1985. Measurement and evaluation in teaching (6th ed.). New York: MacMillan.

Joshua, M.T., 1998. Role of test measurement and evaluation in counseling.

Nigeria Education Journal, 2 (1), Pp. 13-19.

Kelley, T.L., 1927. Interpretation of educational measurements. New York, NY: Macmillan.

Kothari, C.R., 2004. Research Methodology: Methods and Techniques. 2nd Edition, New Age International Publishers, New Delhi.

Mathison, S., 2005. Encyclopedia of Evaluation. Thousand Oaks, CA: Sage

Monday, T.J., 2005. Fundamentals of Test and Measurement in Education. Ultimate Index Book Publishers Press. Calabar, Nigeria.

Nunnally, J.C., 1972. Educational measurement and evaluation (2nd ed.). New York: McGraw-Hill.

Pallant, J., 2011. SPSS survival manual: A step-by-step guide to data analysis using the SPSS program. 4th Edition, Allen & Unwin, Berkshire.

Robson, C., 2011. The Study of the Real World, pp. 114-145. In Stavropoulos, C. (Ed.), Athens, Greece: Gutenberg Publications.