VOL. 11 NO. 6,
FEBRUARY, 2026
E-ISSN 3027-1401
P-ISSN 3026-9148



African Scholar Publications & Research International
www.africanscholarpub.com

Journal of Arts and Sociological Research (JASR)

Measurement Scales in Human and Social Sciences: Nature, Procedure for Development and Validation

Dr. Amos Abiodun Adeyemo (PhD)
Department of Educational Foundations, University of Jos, Nigeria
Corresponding Author: amosadeyemo93@gmail.com
DOI: https://doi.org/10.70382/ajasr.v11i6.0109

Abstract
This work examined the nature of measurement scales, the levels of measurement and the procedure for the development and validation of measurement scales in human and social sciences specifically in the education system. The use of valid and reliable measurement scales in the collection of data in the research process ensure the credibility of research findings. Many researchers collect data from population under study with the use of unvalidated and unreliable research instruments, this affects the usability and applicability of the findings in solving numerous problems confronting the education system in Nigeria. Many other researchers are unaware of the procedure and technicality involved in the development and validation of measurement scales. Through literature review it was established that carefully developed and validated measurement scales ensure objectivity and fairness in data collection, thereby reduce to the barest minimum biasness in testing on one hand and ensure effective generalization and application of findings in the research process. It was recommended that students and by extension researchers within the university system and beyond should be adequately trained on the process for the development and validation of measurement scales.

Key Words: Measurement Scales, Measurement scales development procedure, Levels of measurement, Validity of measurement scales, Reliability of measurement scales

Introduction
Measurement is an integral part of human civilization since inception, and is now being applied in nearly every aspect of human lives and field of endeavours. Every time a test is administered in the school setting, take prescribed doze of medicine, check weather report, cook food, make a financial transaction or interact with people and our environment we embark on a form of measurement. The concept of measurement

is generally associated with precision, accuracy, objectivity, fairness and trustworthiness (Briggs et al., 2023). Since inception the of social and human sciences in the late 19th century discussion on measurement has featured prominently. Specifically, in education, measurement has often been employed, with varying degree of formality, in relation to testing and assessment of psychological constructs (Harris, 2018). This is most obvious when the results of tests or psychological constructs are reported in quantitative terms, and when the differences between test scores or data are interpreted in terms of magnitudes, implying meaningfulness of the units of the scales on which the results are reported (Mari et al., 2023).

Measurement in education includes the development of instruments for obtaining information, procedures for analyzing and evaluating the quality of the information gained from the use of instruments, and strategies for communicating the resulting information to diverse audiences, such as educators, policymakers, parents, and students (Kendeou, 2020). The teaching and learning process is incomplete without the assessment of the extent to which learning has taken place in the learners. Measurement is an integral part of the assessment process as it provides the basis for the teacher and other stakeholders in the education sector to make value judgement about the skills developed and knowledge acquired by learners in the school settings. The development and use of measurement scales play crucial roles in every research process. The research process in the education sector requires the collection of data from the students, teachers, parents and other stakeholders in the education sector with the aim of prescribing solutions to numerous problems confronting effective teaching and learning on one hand and the society on the other hand (Thamarasseri et al., 2018). The purpose of the research embarked upon determines the type and nature of the measurement scale to be developed and validated for use. Some of the measurement instruments use for data collection in social and human sciences include rating scales, questionnaires, interview schedules, check list, test among others. To ensure that these instruments measure what they are designed for and provide a consistent result overtime, various procedure and standards must be followed to ensure their validity and reliability. This will ensure that quality data are collected and meaningful solutions are provided in ameliorating the diverse challenges confronting the education sector.

Statement of the Problem

Many researchers including students across tertiary institutions in Nigeria collect data from samples with the use of unvalidated and unreliable research instruments, this affects the usability and applicability of the findings in solving numerous problems confronting the education system in Nigeria (Adeyemo et al., 2024). In the same vein, many other researchers are unaware of the procedure and technicality involved in the development and validation of measurement scales. These no doubt affect the credibility of research efforts carried out in our tertiary institutions across the nation.

Objectives of the Study

The objectives of the study include to:

i. describe the nature of measurement scales in human and social sciences;
ii. review literatures on the development and validation of measurement scales and
iii. establish the procedure for the development and validation of measurement scales for research purposes.

LITERATURE REVIEW

Measurement is the assignment of numerals to objects according to rules. It is the amount of something, which has been counted, measured or assessed. Measurement is the process of ascertaining the quantity or extent of something (Clauser and Bunch, 2021). The set of numbers used depends upon the nature of the characteristics being measured and the type of measuring instrument used. The set of objects or individuals is determined by the purpose for which the measurements are to be made. The requirements in measurement are a set of objects, a set of numbers, and a rule or rules for the assignment of a number to each object.

Briggs et al., (2023) categorised measurement into physical and mental measurement. Physical measurement is concerned primarily with the dimensions like age, weight, length and capacity among others. These measures are quantitative and therefore require units like years and months, kilogrammes and grammes, centimetres and metres, litres etc. (Harris, 2018). It is objective and does not involve any opinion or subjectivity. On the other hand, mental measurement is the measurement of *constructs*, intangible and unobservable qualities, attributes, and traits that cause certain observable behavior or responses (Thamarasseri et al., 2018). This is what is basically done in the education sector. Thus, Harris (2018) posited that mental measurement is qualitative unlike physical measurement which is quantitative. It is not precise but subjective and indefinite. Kendeou (2020) described some of the basic features of mental measurement. Mental measurements have no specific units rather the units are derived. In any test for instance, a score of 65 has little meaning unless the scores of all test takers are known. Also, the standard adopted by a person is subjective, that is based on personal experience. Furthermore, there is no zero point in mental measurement. Measurement is relative to some standard or norm. For example, an IQ of 80 is said to be below average when compared with the standard value of 100 IQ. Measurement is indirect: for instance, intelligence is measured by what it performs or by a person's performance on various tests and not by physical look.

Furthermore, Lewis and Cook (2020) described two methods of measuring cognitive and non-cognitive abilities in education and psychology; these are Dichotomous scoring and Polytomous scoring. Dichotomous scoring refers to the assignment of one of two possible values based on a person's performance or response to an item. A simple example is the use of correct and incorrect to score a cognitive item response. Alternative-response items are usually scored dichotomously and most cognitive tests involve at least some dichotomously scored items (Mari et al., 2023). Dichotomous scoring is also used in affective measures, such as attitude surveys, behavior checklists, and personality tests. The most common example is scoring that require a response of either yes or no answer. These usually consist of statements written to capture some feature of a construct, such as

adventurousness, and individuals are expected to indicate whether the statements are characteristic of them. On the other hand, polytomous scoring simply refers to the assignment of three or more possible values for a given test item. In cognitive testing, a simple example is the use of rating scales to score written responses such as in essays. In this case, the score values may describe the extent of correctness of a response with differing levels of correctness. For instance, three-point values scale of incorrect, partially correct, and fully correct could be used (Harris, 2018). Such rating scales measure multiple levels of agreement, for instance, disagree to agree or preference from dislike to like. In this case, individuals provide their own responses, which reduce subjectivity in scoring that occurs with dichotomous scoring.

The scale is a continuum used for measuring the level of attributes possessed by an individual or object. It refers to an instrument with a range of point-values used for measuring the set of all the different levels of something from the lowest to the highest. It consists a series of units, each representing an exactly defined position of or degree of presence in the dimension being measured (Briggs, 2021a). In the same vein, Mari et al., (2023) referred to a scale as a continuum that describes the subject's responses to a statement. The different types of scales used in measurement include the Likert scale, semantic differential, check list and ranked items. Likert scale is a five-point scale whereby the middle options provide neutral response (Briggs, 2021a). The options in the Likert scale require the subject to determine the extent to which the respondents agree with a statement. The scale has strongly agreed, agree, undecided, disagree and strongly disagree ratings with undecided as the neutral decision. Semantic differential scale on the other hand has response options that reflect a continuum of bipolar adjectives related to some aspects of the trait being measured (Thamarasseri et al., 2018). For instance, the scale could be:

Fair: __ __ __ __: Unfair Interesting: __ __ __ __: Boring

Aspects of the traits being measured are usually stated in a few words (e.g. My lecturer is …...; the lecture note is …) in this case, an individual's responses position him/her on a continuum (Briggs, 2021a). The responses of "fair" and "interesting" to the statement "My lecturer is …." indicate a positive attitude while the responses "unfair" and "boring" to the statement "My lecturer is …." indicate a negative attitude. Furthermore, in Checklists responses require subjects to identify specific options from which they choose those options that appeal to them (Mari et al., 2023). The options could be yes or no, or Available and Not Available. For Ranked items, the responses in ranked items require students to place a limited number of items into sequential order (Gulturk, 2024). For instance, the ranking could be in the form of 1, 2, 3, 4. Where 1 signifies excellent, 2 very good, 3 good and 4 weak.

Measurement scale therefore, can be referred to as different set of tools used for the collection of data in the research process. They provide the basis for ascertaining the extent to which an individual possessed certain quality of interest in the research

process. They also enable a researcher to measure constructs by assigning numbers to attributes (Mari et al., 2023). Examples of measurement scale include questionnaires, interview schedule, rating scale, checklist among others. In the same vein, scale of measurement refers to an ordered reference standard that provides the basis for comparison; a reference point against which other things can be evaluated (Gulturk, 2024). Also, Briggs (2021a) defined measurement scale as an established set of rules for assigning numerical scores to any directly or indirectly observed variable in order to show the quality or the quantity of the variable. For measurement scale to be valid, it must possess the following characteristics; it must provide for the assignment of numbers to represent all potentially observable values of the variable and it must provide for exclusiveness in the assignment of numbers to the observed quality or quantity of the variable

The Levels of Measurement Scale

Measurement of a particular attribute in an individual or a thing can be done on one of four different scales of measurement. These are the nominal scale, ordinal scale, interval scale and ratio scale. These scales are hierarchical in structure. The nominal scale is the lowest scale while the ratio scale is the highest ordered of them. Nominal Scale is also known as classificatory scale. It represents the lowest level of refinement. It involves the use of labels or categories, which are strictly for referencing. Numbers or other symbols are used simply to classify an object, person, and characteristic or to identify the groups to which various objects belong (Gulturk, 2024). For example, classification of persons into males and females, rural and urban categories, number on the jerseys of members of a football team, the number on library books among others. The numbers assigned do not indicate any order. They are used only to name, identify or classify. Two assumptions are made regarding the assignment of numbers in nominal scales (Briggs, 2021b). These assumptions are that a different number is assigned for each quality or quantity being measured and that there exist a number for every observed or potential quality or quantity.

On the other hand, ordinal scale is also known as ranking scale and the second lowest ordered scale of measurement. It recognizes and allows for ordering or ranking of the individuals with reference to the trait or the attributes being measured (Briggs, 2021a). According to Gulturk (2024) in this scale, each of a set of objects is assigned based on certain rules, to one of a set of ordered categories. These categories differ quantitatively from one another. For instance, pupils in a class can be categorized into above average, average or below average based on reading ability. In such a classification, each class is higher than the classes below it, and lower than the classes above it. The numbers used in identifying observations are called ranks. Briggs, Maul, and McGrane (2023) gave three assumptions in the assignment of numbers in ordinal scale; these are that a different number is assigned for each quality, that there exist a number for every observed or potential quality and that the numbers represent differences in some magnitude among the aspects of the variable being measured. Interval scale is the third

scale of measurement. It tells or shows the difference between two objects or persons being measured. It possesses all the qualities of the nominal and ordinal scales, that is, it allows for classification or referencing, and shows order or rank (Gulturk, 2024). Additionally, it answers such questions as "by how much"? "how far apart"? etc. with respect to the attribute being measured. This means that the law of equal intervals is preserved on measurements that are taken on the interval scale. However, it does not have a true (or absolute) zero-point (Kendeou, 2020). By this, it means that although a zero point might be on the scale and all other points located with reference to the zero point, the measure at the point is not zero in the real sense. The zero point of the interval scale can be placed arbitrarily and does not indicate absence of the property measured. Some examples of data collected on this scale include scores of students in an achievement test; data collected from an administered questionnaire. However, Gulturk (2024) made the following assumptions about interval scales: that a different number is assigned to each quality, that there exists a number for every observed or potential quality, that the numbers are assigned to represent differences in some magnitude among the aspects, that the numbers assigned represent equal units of counting. Ratio scale is the most elaborate, highly ordered, scale of measurement. It has all the properties of the interval scale. That is, it can be used for classification, it shows ranking (order) and it has equal intervals (Haruna, 2018). Additionally, it has a true (absolute) zero position on the scale. This point actually implies a complete absence of the characteristics being measured on the individual. The instruments used for measuring physical quantities (in Biology, Physics, Chemistry etc.) are operating on this level (Briggs, 2021b). Assumptions are made regarding the assignment of numbers in ratio scales. These are; that a different number is assigned for each quality; that there exists a number for every observed or potential quality; that the numbers represent differences in some magnitude among the aspects of the variable; that the numbers assigned represent equal units of counting and that the starting point for the scale represents a true zero point (Clauser & Bunch 2021).

Basic Qualities of Measurement Scales

In order to ensure the credibility of a measurement scale developed to assess or collect data on any psychological construct, Clauser and Bunch (2021) opined that such measurement instrument must fulfill specific conditions. Kendeou (2020) stated that evaluation technique is judged for its adequacy, efficiency, consistency and suitability as a measuring device based on commonly accepted qualities. Some of these qualities are validity, reliability, objectivity, adequacy, usability and discrimination.

Validity is the watchword upon which the entire superstructure of any measurement scale is based. It is the extent to which an evaluation technique accurately measures what it is intended to measure (Lewis & Cook 2020), The three types of validity used in educational and psychological measurements as identified by Cizek (2020) are content, criterion-related and construct validity.

Content validity represents the extent to which a scale measures a representative sample of the subject matter. It is determined based on experts' judgment (Harris, 2018). The developer then

examines carefully the boundaries of the contents and objectives associated with the units. The overall judgment about content validity is based on the extent of agreement among experts of the adequate representation of the items of scale being developed for measurement of abilities or psychological constructs of the universe of content of a subject matter. (Briggs et al., 2023). Criterion-related validity is defined as the extent to which a person's performance on a test is related to some other valued measure of performance (Cizek, 2020). It is based on relationship between scores on the test and appropriate external criterion. Criterion related validity according to Ansari and Sheema (2023) can be concurrent or predictive. Concurrent involves the use of criterion measurements obtained appropriately at the same time as that of the test or scale to be validated. It is considered when using a test or a measuring instrument to distinguish between two or more groups of individuals whose status at the time of data collection is different (Harris, 2018). Predictive validity on the other hand is used to predict the future status of an individual for a particular trait being measured. Construct validity is the accuracy with which an evaluation technique designates an individual in terms of a stated psychological trait. It is the extent to which test performance can be interpreted in terms of certain psychological constructs (Cizek, 2020). It helps to determine psychological quality or trait such as intelligence, creativity, anxiety that is assumed to exist in order to explain some aspects of behavior. Construct validity of instruments can be established through three different methods: correlation method, group difference method and factor analysis method. Correlation method involves correlating the scores on an instrument with scores on the other measures of some behavior (Harris, 2018). Group difference method involves comparison of test performance of a group that exhibits the trait being measured with that of another group that does not possess the trait. Factor analysis on the other hand involves reducing a large number of measures to smaller number called factors by discovering which items go together or measure the same thing. (Briggs, 2021b). According to Glen (2021) factor analysis is an attempt to bring inter-correlated variables together under a more general fundamental variable.

Lewis and Cook (2020) identified two approaches to finding the validity of a measurement instruments to include logical validity and empirical validity. Logical validity involves using logical analysis to judge precisely what an instrument measure. It is based on consistency of thought and can be found by two methods; deductive and inductive validity. In deductive validity, deduction proceeds from general judgment to particular judgment. It finds out if items on a scale correspond to the definition of the trait being measured and the trait has to be objectively defined. For inductive validity judgment proceeds from particular to general. Assignment of name to the trait on the basis of the particular trait and then the validity is found. Instead of seeing whether a test measures a particular trait, what it measured is discovered and the scale is declared valid for that trait. On the other hand, empirical Validity is also known as statistical validity. Here the validity for a scale is determined by correlating it with a criterion. Hence, the objective of a course or training could form a good criterion.

Reliability is the degree of consistency with which a test measures what it purports to measure accurately (Cizek, 2020). The factors that affect an evaluation technique reliability as identified in literature (Harris, 2018; Cizek, 2020) include length of the instrument (the scale

should be long enough to be valid and short enough to be useful), time limit (the longer the time apportioned for the scale the better the reliability estimate), group heterogeneity (the wider the range of abilities of the test takers the higher the estimate reliability), testing condition (unsystematic factors such as light, ventilation, noise cause variations in test scores of candidates) and objectivities in test administration (errors in test administration and test scoring such as halo effect, generosity and penalty error that occur due to lack of objectivity in testing lowers the reliability of the test). The following methods identified by Lewis and Cook (2020) and Cizek (2020) can be used to estimate the reliability of measurement instruments in educational research. Test-retest reliability or co-efficient of stability in which an identical test is administered to the same group of students on different occasions and the scores correlated. Parallel form or equivalent method in which two equivalent tests of different contents but measuring the same construct are given to the same group of students on different occasions and the scores are correlated. The split-half method is a measure of internal consistency in which a single test is administered but each individual score on each item is split into odd-and-even numbered scores, which are then correlated. Kuder-Richardson method measures the internal consistency of dichotomously scored test from a single administration without splitting to test the homogeneity of the test items. The method uses two formulae K-R 20 and K-R 21 respectively. Coefficient Alpha method developed by Cronbach estimates the reliability of instruments that are scored dichotomously and those not scored dichotomously. It is computed from the variance of individual test items and variances of the total test score.

Another basic quality of measurement scale is objectivity. According to Lewis, and Cook (2020) objectivity is the degree to which a measurement instrument is uninfluenced by the belief of its user. It is the extent to which test scorers obtain the same results on a test so as to reduce scorer biasness. The degree of objectivity of educational instrument is determined by analyzing the testing and scoring procedures as well as discrepancies in experts' judgment of the test. Finally, test fairness deals with appropriate use of scale responses and in a social term that represents someone's value judgment. A fair test is one that yields comparably valid inferences from person to person and group to group (Cizek, 2020). A fair test is one that affords all respondents equal opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the test/scale purpose (Mari et al., 2023).

Procedure for the Development and Validation of Measurement Scales

Boateng et al. (2018), Carpenter (2018) and Lamm et al. (2020) described the process for the development of measurement scales. Boateng et al. (2018) identified three major phases that span nine steps for the development and validation of measurement scales. Carpenter (2018) identified 10 steps in scale development and reporting while Lamm et al. (2020) describe a robust five steps in scale development for empirical studies. The works of Carpenter (2018) and Lamm (2020) were all captures in that of Boateng. According to Boateng (2018) the phases of scale development are item development phase which comprises step one and step

two, scale development phase comprises step three to step six and scale evaluation phase which is made up of step seven to step nine.

Phase 1: Item development

Step 1: Identification of the domains and item generation

Identification of the domain: This entails articulating the domain to be measured. A domain or construct refers to the concept, attribute, or unobserved behavior that is the target of the study. Therefore, the domain being examined should be decided upon and defined before any item activity (Boateng et al., 2018).

Item generation: This process is also called question development or item generation. There are two ways to identify appropriate questions: deductive and inductive methods. The deductive method is based on the description of the relevant domain and the identification of items. This can be done through literature review and assessment of existing scales and indicators of that domain while the inductive method involves the generation of items from the responses of individuals (Carpenter, 2018). Qualitative data obtained through direct observations and exploratory research methodologies, such as focus groups and individual interviews, can be used to inductively identify domain items.

Step 2: Assessment of Content Validity.

After the generation of the items of the scale, the content validity is determined. This entails establishing the adequacy with which a measure assesses the domain of interest. It is done to ensure content relevance and content representations of the domain of the measurement scale. This is mainly assessed through evaluation by experts and target population judges (Lamm et al., 2020).

Phase 2: Scale Development

Step 3: Pre-testing Questions

After the establishment of the content validity, the scale is pretested. This entails administering the scale to a similar group before the main study. The group for pre-testing is expected to possess similar characteristics with the group for the main study. This help to minimizes misunderstanding and subsequent measurement error (Boateng et al., 2018).

Step 4: Sampling and Survey Administration

This involves ensuring adequate sample size for data collection. It is suggested that a minimum of 300–450 elements are required to observe an acceptable comparability of patterns, and that replication is required if the sample size is < 300 (Boateng et al., 2018). Survey administration can be done using paper and pen/pencil interviewing (PAPI) or Computer Assisted Personal Interviewing (CAPI) on devices like laptops, tablets, or phones to collect data.

Step 5: Item reduction analysis

Item reduction analysis is conducted to ensure that only parsimonious, functional, and internally consistent items are ultimately included in the final scale. Therefore, the goal of this phase is to identify items that are not or are the least-related to the domain under study for deletion or modification. This is done after the conduct of factor analysis. Items with factor loading of 0.4 and above are retained while items with less than 0.4 are dropped (Carpenter, 2018).

Step 6: Extraction Factors

Factor extraction is the phase in which the optimal number of factors, sometimes called domains, that fit a set of items are determined. This is done using factor analysis. A number of analytical processes have been used to determine the number of factors to retain from a list of items, commonly available methods to determine the number of factors to retain include a scree plot, the variance explained by the factor model, and the pattern of factor loadings (Lamm et al., 2020).

Phase 3: Scale Evaluation

Step 7: Test of Dimensionality

The test of dimensionality is a test in which the hypothesized factors or factor structure extracted from a previous model is tested at a different time point in a longitudinal study or on a new sample. Tests of dimensionality determine whether the measurement of items, their factors, and function are the same across two independent samples or within the same sample at different time points (Boateng et al., 2018). Such tests can be conducted using independent cluster model (ICM), confirmatory factor analysis, bi-factor modeling, or measurement invariance

Step 8: Test of Reliability

Reliability is the degree of consistency exhibited when a measurement is repeated under identical conditions. A number of standard statistics have been developed to assess reliability of a scale (Carpenter, 2018). Cronbach's alpha or test–retest reliability is predominantly used to assess reliability of measurement scales.

Step 9: Test of Validity

Validity is the extent to which an instrument measures the latent dimension or constructs it was developed to evaluate. Validation is an ongoing process that starts with the identification and definition of the domain of study and continues to evaluating necessary data with which hypotheses involving the nature and meaning of a variable are tested. The validity of an instrument can be examined in numerous ways; the most common are content validity, criterion and construct validity (Boateng et al., 2018).

Review of Empirical Studies on development and validation of Measurement Scales

Tentama and Nabilah (2020) carried out a study on development and confirmatory factor analysis of a measurement scale in Yogyakart. The aim of the study was to develop and test the validity and reliability of the employability skills scale. The study also aimed to establish the dimensions of the construct of the scale using factor analysis. The scale was used to collect data from a sample of 90 students of class VII at SMK X in Yogakarta. After the factor analysis three factors were extracted and these are career identity, personal adaptability and social and human capital. The result of the analysis showed that the dimensions of the scale were valid and reliable.

Also, Caballero et al. (2020) carried out studies on developing and validating measurement scales to promote the employability capital of university students in Australia. The work developed and validated a set of instruments for measuring the factors of employability capital of university students. Confirmatory factor analysis was carried out to determine the factor structures of the scales. Three studies based on three samples of 816, 1,082, and 1,088 students were used for the studies. The content validity, dimensionality, and reliability, as well as convergent, discriminant, and nomological validity of the scales were established. The results of the studies showed that the scales were valid and reliable and can be used to measure employability constructs among university students.

Furthermore, Farenanda et al. (2021) carried out a study on the construction of work readiness measurement scale for fresh graduate in Malaysia. The study developed and validated a work readiness measurement scale for fresh graduates. The initial scale was made up of 72 items, however after the conduct of factor analysis the final scale evolved with 24 items as 48 items were judged to be complicated by loading on more than one factors or had an eigen value less than 1 were eliminated. The reliability coefficient of the scale was 0.93 while the content validity index was 0.84. It was concluded that the work readiness scale for fresh graduates had good psychometric properties and can be used for assessment and research.

Also, Bennet and Ananthram (2022) conducted a study on the development, validation and deployment of measurement scale among higher education students in Australia. The aim of the study was to provide higher education students a measurement scale which can be used to self-assess their perceived employability skills. A Perceived Employability Measurement Scale (PES) was developed and piloted with a sample of 1849 university students. The final scale was administered with students from over 15 Australian universities, resulting in a total sample of 10,193 students. The scale was validated using exploratory and confirmatory factor analyses, resulting in a 14-factor employability measure that met the acceptable validity thresholds. The multi-factor perceived employability scale validates the USEM model and had implications for theory and practice.

CONCLUSION

Close examination of literature review shows that most researchers used factor analysis to determine the construct validity of measurement scales while Cronbach Alpha method of reliability was used to establish the internal consistence of the items of the measurement scales.

However, there are some inadequacies in the process involved in the development and validation of measurement scales by some researchers. For example, some of the studies did not estimate the reliability coefficients of some of the measurement scales developed, some did not establish the construct validity of the scales developed and used for their studies. Also, most of the studies reviewed did not determine the differential item functioning of the items of the scales developed and this call to question the issue of fairness in terms of gender, location, socio economic factors among other variables measured given the dynamics of the sampled population. Perhaps, validity, reliability and fairness are essential characteristics of credible research instruments.

RECOMMENDATION

Based on the findings of the study, the following recommendations were made:

- All researchers should adequately undergo training on the process and procedure involved in the development and validation of measurement scales; this will ensure that all research instruments used for data collection are valid, reliable and fair thereby enhancing the credibility of research findings and as well ensure applicability beyond the immediate locations of which research efforts were carried out.
- It was also recommended that researchers are to ensure the elimination of biasness in items of measurement scales through the establishment of differential item functioning for every measurement scale developed and used in the research process.
- It was equally recommended that researchers should be proper oriented on the criteria for the adaptation or adoption of research instruments in the research process.

REFERENCES

Adeyemo, A. A., Dadughun, S. I. & Wakjissa, S. G. (2024). Development and validation of employability readiness scale for university undergraduates in North-Central Nigeria. *IOSR Journal of Research and Method in Education. 14*(1), 43-50. Available Online: http://doi.org/10.9790/7388-1401014350

Ansari, M. M., & Sheema, K. (2023). An in-depth examination of validity assessment: Exploring diverse methodologies and dimensions of validity in social research studies. *Asian Journal of Agricultural Extension, Economics & Sociology* 41 (10):772-82. https://doi.org/10.9734/ajaees/2023/v41i102224.

Boateng, G. O., Neilands, T. O., Frongillo, e. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social and behavioural research: A primer. *Front Public Health, 6,* 149-169.

Briggs, D. C; Maul, A., & McGrane, J. (2023). *On the nature of measurement.* Retrieved on the 15th February, 2026 from https://scispace.com/pdf/on-the-nature-of-measurement-2d1f433r5f.pdf

Briggs, D. C. (2021a). *Scaling and its relationship to measurement*. In B. Clauser (Ed.) History of Educational Measurement. Routledge.

Briggs, D. C. (2021b). *Historical and conceptual foundations of measurement in the Human Sciences: Credos and Controversies*. Routledge.

Caballero, G., Alvarez-Gonzalez, P. & Lopez-Miguens, M. J. (2020). How to promote the employability capital of university students. Developing and validating scales. *Studies in Higher Education, 45*(12), 2634-2652.

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researcher. *Communication Methods and Measures, 12*(1), 25-44. Doi:10.1080/19312458.2017.1396583

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Clauser, B. & Bunch, M. B. (2021). *The History of Educational Measurement.* New York: Routledge.

Farenanda, K. W., Afifudin, I., Anggi, N. A., Putri, U. K., Farica, A. N., Shavira, I. P., Hemalia, A. & Aftina, N. H. (2021). T*he construction of work readiness scale.* Retrieved on August 1, 2023 from http://repository.urecol.org/index.php/proceeding/article/view/1774

Glen, S. (2021). *Factor Analysis: Easy Definition.* Retrieved February 25, 2023 from https://www.statisticshowto.com/factor-analysis/

Gulturk E. A. (2024). Scale adaptation and redevelopment: A review on validity and reliability. *Journal of Cellular Molecular and Immunology, 3(1):26-32*.

Harris, D. (2018). *Educational Measurement: Issues and Practice.* Retrieved February 15, 2023 from https://www.wiley.com/en-us/Education+Measurment%3A+issues.

Haruna, A. S. (2018). *Educational Research*. SK Amodu printing and publishing house.

Kendeou, P. (2020). The assessment of reading for understanding. In P. D. Pearson, A. Palincsar, S., G. Biancarosa, & A. Berman (Eds.). Reaping the Rewards of the Reading for Understanding Initiative. *Washington, DC: National Academy of Education, 67-142*.

Lamm, K. W., Lamm, A. J., & Edgar, D. (2020). Scale development and validation: methodology and recommendation. *Journal of International Agricultural and Extension Education, 27*(2), 24-35. https://doi:10.5191/jiaee.2020.27224.

Lewis, D. & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with Contemporary assessment design principles. *Educational Measurement: Issues and Practice, 39: 8-21.* https://doi.org/10.1111/emip.12318

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement, 2nd edition.* Cham: Springer.

Tentama, F., & Nabilah, B. R. (2020). Development and confirmatory factor analysis of employability skill scale. *International Journal of Scientific and technology research, 9*(2), 538-542.

Thamarasseri, I., Bhat, M. S., & Sheikh, M. (2018). *Educational measurement and evaluation*. APH Publishing Corporation.