British Journal of Education, 14 (8),1-16, 2026
Online ISSN: 2054-636X (Online)
Print ISSN: 2054-6351(Print)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development-UK

The Concept of Scale in Educational Measurement

Bulejo Olajide Olanrewaju
Institute of Education, Faculty of Education, Ekiti State University, Ado – Ekiti, Nigeria
ORCID ID: 0000-0003-1571-5431

Rhoda Oluwaseun Kolawole
Department of Educational Psychology, School of Education, Federal College of Education, Ilawe Ekiti

doi: https://doi.org/10.37745/bje.2013/vol14n8116 Published August 23, 2026

Citation: Olanrewaju B.O. and Kolawole R.O. (2026) The Concept of Scale in Educational Measurement, *British Journal of Education*, 14 (8),1-16

Abstract: *The study is a resume of the concept of scale in educational measurement. Rating scale was defined as a technique through which the observer or rater categorizes the objects, events or persons on a continuum represented by a series of continuous numerals. Attempt was made to clarify the purpose of rating scales to include measuring teachers' performance/effectiveness and personality, anxiety, stress, emotional intelligence and so on. The uses of rating scales to include helpful in writing reports to parents, helpful in filling out admission blanks for colleges and helpful in finding out student needs among others. Types of rating scales includes comparative rating scale, graphic scale and numeric scale. Scaling affective characteristics was seen as having the thorough review of the theory concerned with the affective characteristics and to generate the perceptions, attributes or behaviors of a person with high or low levels of this characteristic. Stages involved in the development of rating scale was enumerated to include; defining the construct to be measured, designing the scale, generating an item pool, page layout, administering the scale, sample size, checking the data, coefficient – alpha and factor analysis.*

Keywords: rating scale, measurement, construct, validity, factor analysis

INTRODUCTION

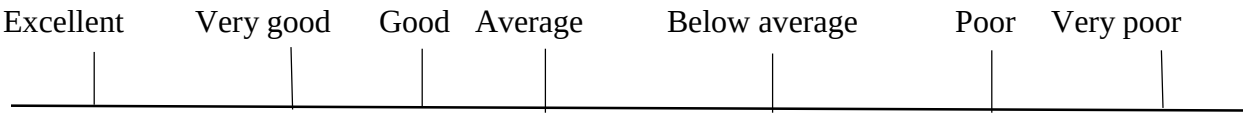
A Rating scale is defined as a technique through which the observer or rater categorizes the objects, events or persons on a continuum represented by a series of continuous numerals. The experience may be direct or indirect or remembered. A rating scale is one of the many instruments used in measuring affective behavior. It attempts to ascertain the degree of intensity of a variable. It usually has two, three, five, seven, nine or eleven points on a line with descriptive categories at both ends

followed sometimes with a descriptive category in the middle of the continuum, too. (archive.mu.ac.in)

A rating scale according to Olanrewaju (2019) is an assessment instrument used to judge or rate the quality of a particular trait, characteristic, or attribute of the pupil based on pre-determined criteria. Teachers can use rating scales to record observations and students can use them as self-assessment tools. Teaching students to use descriptive words, such as **always**, **usually**, **sometimes** and **never** helps them pinpoint specific strengths and needs. Rating scales also give students information for setting goals and improving performance. In a rating scale, the descriptive word is more important than the related number. The more precise and descriptive the words for each scale point, the more reliable the tool.

Rating scales can be used to assess the individual once only or on repeated occasions to show developments over time. They can be completed using information derived from direct observations or from existing information contained in running records or anecdotal records (Cohen, et al 2007). They require relatively little time to complete especially if information is already available and direct observations are not required. If they are designed well, they provide clear, objective criteria against which ratings are made. They can be used to provide large amounts of information which can be analyzed to look for patterns and they can be used to present the information to parents and others. It is a very useful device in assessing quality, especially when quality is difficult to measure objectively. For Example, 'How good was the performance?' is a question which can hardly be answered objectively.

Rating scales record judgment or opinions and indicates the degree or number of different degrees of quality which are arranged along a line in the scale. For **e**xample: How good was the performance?



This is the most commonly used instrument for making appraisals. It has a large variety of forms and uses. Typically, they direct attention to a number of aspects or traits of the thing to be rated and provide a scale for assigning values to each of the aspects selected. They try to measure the nature or degree of certain aspects or characteristics of a person or phenomenon through the use of a series of numbers, qualitative terms or verbal descriptions.

Purpose of rating scales:
Rating scales according to Olanrewaju (2019) have been conceptualized for the purpose of measuring the following:

i. Teacher Performance/Effectiveness
ii. Personality, anxiety, stress, emotional intelligence etc.
iii. School appraisal including appraisal of courses, practices and programmes.

British Journal of Education, 14 (8),1-16, 2026
Online ISSN: 2054-636X (Online)
Print ISSN: 2054-6351(Print)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development-UK

Uses of rating scales

They are of great advantages in the following fields like:

i. Helpful in writing reports to parents
ii. Helpful in filling out admission blanks for colleges
iii. Helpful in finding out student needs
iv. Making recommendations to employers.
v. Supplementing other sources of understanding about the child
vi. Stimulating effect upon the individuals who are rated.

Types of rating scales

Rating scale has among others as its common types

i. comparative rating scale
ii. graphic scale
iii. numeric scale
iv. semantic differential scale
v. rank/order scale
vi. likert summated rating scale

Why scale in the first place? Or, stated differently, if we somehow stopped all scaling tomorrow, would that be any great loss? According to Robert (1996), the first thing was that it is important when making decisions about allocating people to tasks that we have some methods by which to make good allocations…… and to do this in an efficient manner. We tossed out several examples such as, if you were going to have your gall bladder out, heaven forbid, it would be nice to have some assurance that the person who was going to do it had training and could perform the task with minimal risk to you, etc.

Another reason was to make diagnoses of a personal nature…… to better known oneself…….regardless of whether or not that impacts on decisions on what you might do. For example, scales can provide a person with personal information that might enable him or her to change what he or she does.

A third reason was to provide some baseline information about things. For example, what if we had a task that needed to be done and, we had decided or come to the conclusion that it would take X amount of skill in order to perform the task, a scale might provide baseline information about whether anyone had sufficient skill to carry out the task…..or the baseline might simply provide information about how many might be available in a pool that could do the task if we needed it done.

In general, the goals of psychological scaling are to create a meaningful and valid scale, to locate persons (or objects) in their correct relative positions and to arrive at distances amongst persons (or objects) that reflects true differences amongst them. A scaling model is a plan to develop a scale on which to place people or objects. For example, a scale could be, permissiveness, intelligence and introversion.

There are several basic and universal steps involved in a scaling; the foremost of all is the "scale concept". What is it that you are trying to scale people or objects on? Intelligence, Dogmatism, Attitudes about statistics. Second, what is the "target" of the scaling? People or objects? Third, what is the plan or method by which the scaling will take place? Is it people, direct questioning, observation, or using calibrated stimulus item? Fourth, what instrument and/or collection of stimulus items will be used? Fifth, what is the mechanism by which scales values or scores for people (or objects) be determined and assigned? And finally, given the purpose of the scaling, how does one determine if the scaling model is implemented in a reliable and valid way? Does this scaling model or plan accomplish what it is supposed to do? Well, the above were some initial things to worry about (Robert, 1996).

However, the worries highlighted by Robert above were put to rest, taking into consideration the work of psychologists and tests experts on scaling and educational measurement. The work of Cronbach, (1971). Test Validation, Alonge, (2004). Measurement and Evaluation in Education and Psychology, Bandele, (1994). A Monograph on Construction of Questionnaires and Rating Scales, Omirin, (1999). Construction and Validation of Science-oriented Attitudinal Scale, Adebule, (2002). Development and Validation of an Anxiety Rating Scale in Mathematics for Nigerian Secondary Schools and Kolawole, (2005). Assessment and Measurement in Education and host of others have added credence to scaling and instrument developments.

According to Adebule (2002), measurement always takes place in more or less complex situations in which innumerable factors may affect both the characteristics being measured and the process of measurement itself. Similarly, the measurement of any psychological or social characteristics presupposes a constant set of known conditions among the factors relevant to it and to the process of measuring it. Alonge, (2004) defined measurement as the process of describing with numbers according to rules, the quantity, quality or value of a product, performance or understanding. The set of numbers depends on the nature of the characteristics being measured and upon the type of instrument used. A measurement procedure therefore consists of a technique for collecting data plus a set of rules for using these data. To be useful, the data collection, techniques and the rules for using the data must produce information that is not only relevant but correct.

In the study of "Basic Technique in Practical Research" by Omirin (1999), the most common mistake made by beginning researcher is to use wrong statistical technique for data analysis. This mistake often arises from lack of proper understanding that the natures of all measurement are not the same. Measurement obtained by measuring the weight of students in a class is different from those obtained by measuring their career interests using vocational interest inventory. The precision with which we can measure a student's interest in mathematics is not as high as that with which we can measure the height to the nearest millimeter, (Omirin 1999)

In the same vein Robert (1996) noted that a measurement scale is a set of rules for quantifying or assigning numbers to a particular characteristics or variable. When scaling, there are at least 3 dimensions that come into play: people or objects, stimuli and responses. "people" are those who you

want to scale (P1, P2 etc.), by presenting them with various stimuli (S1, S2 e.t.c), and based on that…… people make responses (R1, R2 etc) to the stimuli.

Stages in the Development of Rating Scale

Measurement according to Murphy, (2005) is the process of ascribing numbers to persons or objects in such a way that some attributes of the persons being measured are faithfully reflected by some properties of the numbers. It is the application of numbers according to some specified rules (Alonge, 2004).

A scale is an instrument used in collecting information about the extent of presence of a psychological concept or trait in a respondent. It is a collection of statements which when added or summed together measure some hidden, underlying or convert structure. The items put together do not measures the same face of the trait but they measure similar or related facets of the construct of interest. Each scale of construct possesses certain properties. These properties determine the appropriate use of certain statistical scale of measurement. Scale development is a process of creating a measure that assesses a particular attitude, interest, skill, knowledge and/or ability.

A scaling procedure permits a person to be assigned a numerical score indicative of his/her position on the attitudinal dimension. Researchers commonly describe the different ways they measure things numerically in terms of scales of measurement, which come in four ways: nominal, ordinal, interval or ratio scales.

Nominal scale: According to Adebule (2002), nominal scale is the lowest level of quantification. It is called the naming or classificatory scale. The numbers in this scale is just for classification, categorize or name items into discrete position and no magnitude relationship to one another. For example, a study that borders on gender issues in a school can be classified using M for male and F for female or the classification may be by religion, using A for Christianity, B for Islam and C for paganism. Nominal scale does not possess the attributes or characteristics of magnitude, equal interval or absolute zero.

Ordinal scale*:* The ordinal scale has the merit of rank order which determines the order of relation between things. Omirin (1999) said that ordinal scale is an improvement on the nominal scale, in that it is possible not only to indicate a difference in the characteristics being measured but also the amount or degree of the difference. However, Alonge (2004) revealed that ranking order is deficient in that the distance between ranks could not be explained as interval. For instance, after a test, a teacher may rank the students' relative position as 1^{st}, 2^{nd}, 3^{rd} and so on. We cannot say that the difference between 1^{st} and 2^{nd} is the same as that between 2^{nd} and 3^{rd} position and so on. Scale at ordinal level reflects only magnitude but not interval or absolute zero point, that is to say that ordinal scale has all the characteristics of the nominal scale and in addition, it can be ranked/ordered

Interval scale: interval scale is the third level in the hierarchy of measurement. It has been ascertained that interval scale possesses all the properties of the first two levels in addition to the property of equal units. That is, it has magnitude and quality of equal interval between units of measurement but

no true zero base. Omirin (1999) pointed out that by this property; differences between various levels of the categories on any part of the scale reflect equal difference in the characteristics measured. Here arithmetical computations are possible. For example; addition, subtraction, averages and so on. However, high level statistical manipulation or mathematical computations are not appropriate at this level.

Adebule (2002) gave an example of a set of scores of students in an examination to illustrate this. A set of scores of 10,20,30,40 can be explained to have equal interval of 10. While according to Omirin (1999) in a test, A gets a score of 90 while B gets a score of 85, not only has student A performed better than student B, his performance is five points better. If student C scored 80, then student B would have outperformed student C by as much as student A outperformed student B. a typical interval scale is the thermometer used to measure temperature. There are differences between 5 and 10 degrees as between 40 and 45 degrees on a thermometer. One major weakness of the interval scale is that comparison cannot be made because it has no clear zero level, a zero point is just another point on the scale. It does not reflect the starting point on the scale nor did a total absence of the characteristics being measured. (Omirin, 1999)

Ratio scale: This is the fourth and the last in the hierarchy of measurement. It is referred to as the strongest of the measurement. Ratio scale is a scale of measurement that has all the three characteristics of measurement, magnitude, equal interval and absolute zero point. Here all the properties of nominal, ordinal and interval scales are embedded in the ratio scale. Omirin (1999) noted that in the physical science, the ratio level include instrument like thermometer, meter mile, weight scales etc, they all have clear cut zero levels. For example, the zero point on a weighing balance indicates the complete absence of weight; similarly, no height and 0^{o}c means no temperature etc. Here, at this level of quantification, comparisons of ratio are allowed to be made concerning the attributes being measured, sophisticated manipulations of figures, mathematical models can be carried out.

Most of the data used in educational research are either nominal, ordinal or interval data. The aforementioned scales above falls into these first three scale of quantification, while ratio scales are really used in the behavioral science. To select the appropriate statistical tests for analyzing data, it becomes germane for the researcher to be able to identify the scale in which the data were collected whether nominal, ordinal or interval (Bandele, 1995).

Scaling Affective Characteristics

One thing that is significant of affective scaling is to first have the thorough review of the theory concerned with the affective characteristics. The next step is to generate the perceptions, attributes or behaviors of a person with high or low levels of this characteristic. Daramola (2010) illustrates two similar approaches to this task- namely, the domain reference approach and mapping-sentence approach.

In domain-reference approach to affective scale construction by Daramola (2010), the target and the direction of the affective characteristics are first addressed, and then the intensity aspect is considered. It is proposed that this technique be adapted to also include a statement of the priori, judgmentally-developed categories which the clusters of statements are intended to represent. It was noted that since affective characteristics can be described as having intensity, direction and a target, a frame work was described for developing statements for an affective instrument by carefully sampling from a universe of content. After developing such statements, we typically obtain the responses of selected individual to these statements and claim that we have measured the intensity and direction of their effect towards the particular target object.

Olanrewaju (2014) opined that, measurement begins with the concept of a continuum on which people can be located with respect to some traits or constructs. Instruments in the form of test items are used to generate numbers called measures for each person. It is important to realize that the test items are also located on the continuum with respect to their direction and intensity. Variations of judgmental and empirical techniques are used to scale the items. Some scaling procedures calculated numbers called calibrations, which indicates the location of an item on the underlying affective continuum.

During the measurement process, people are measured by the items that define the trait underlying the continuum. That is, people are located on this abstract continuum at a point which specifies a unit of measurement to be used to make "more or less" comparisons among people and items. According to Olanrewaju (2019), the requirements for measuring affective characteristics are:

1. the reduction of experiences to a one-dimensional abstraction (i.e continuum);
2. more or less comparisons among persons and items;
3. the idea of linear magnitude inherent in positioning objects (i.e people along a line) and
4. a unit determined by a process which can be repeated without modification over the range of the variable. (Olanrewaju, 2019)

This process, which can be repeated without modification is actually the measurement or scaling model used to describe how people and items interact to produce measures for people. Several scaling models have received attention over the last 60years, with particular emphasis during the last 20years. Here the techniques under consideration include Thurstone's (1931) Equal-Appearing Interval technique, Likert's (1932) Summated Rating technique.

Thurstone technique begins with a set of belief statements (i.e attributes) regarding a target object. These statements are then located (i.e calibrated) on the favorable/unfavorable evaluative dimension through a judgmental procedure that results in a scale value for each belief statement. Thurstones' procedure in his early use of paired comparisons after the set of items had been scaled by the judges, items were paired with other items with similar scales values; and sets of paired comparisons were developed. In some cases, each item was paired with all other items from other scales on the instrument and respondents were asked to select the item from the pair that best described the target object.

Researchers are often concerned with the differences among these scales of measurement because of their implications for making decisions about which statistical analyses to use appropriately for each. At times, they are discussed in only three categories: nominal, ordinal and continuous (i.e, interval and ratio are collapsed into one category). The process of scale development has been described by Olanrewaju (2019). The seven essential steps that were identified are adopted in this study. They are as follows:

Step 1: Define the construct to be measured

At this stage, the researcher is expected to identify the construct to be measured and explain the construct through the underlying theories or models proposed for the construct. Netemeyer et al. (2003) explain that it is crucial to assess the boundaries of the construct domain and not to include extraneous factors or domains of other constructs.

Step 2: Design the scale

This stage has to do with the response format to be used in making submission by the respondents. The Likert (1932) format of response is the most popular of all response formats (Hopkins, 1998) used for summated rating scales. However, the Likert scale varies by agreement, evaluation or frequency. The agreement request requires respondents to express the level or degree to which they agree to a statement. The responses are usually expressed as strongly agree, agree, uncertain, disagree and strongly disagree. The evaluation form of response to the Likert scale asks for evaluating rating of each item. The response is expressed as poor, below average, average, above average and excellent. The ratings could be presented in reverse form starting with excellent and ending with poor. The frequency response forms on the Likert scale asks for the rate of occurrence of something. The options usually presented on such format include: rarely, seldom, occasionally, frequently and usually.

Usually, between five and seven response options are adequate for the Likert format scales as too much information is lost with fewer than four options and not much is gained with more than ten. Response options are expected to be odd numbers e.g 5,7 or 9 if the researcher wants to aloe for fence – sitting and even numbers e.g 4,6,8 or 10 if fencing –sitting or being neutral is to be allowed.

Step 3: Generate an item pool

Here, the initial items on the scale are being generated. Items are being generated from the outcome of review of related literature, observation, tests, inventories and conservation with others. At this stage, the researcher is expected to generate a large (e.g more than needed) number of items. This is to ensure that at the end of the validation exercise, there would be enough items left on the scale. The sentence or statement which respondents are expected to agree to, rate, evaluate or express in frequency is the stem and list of alternatives from which the respondents select is the options, (Popham, 2006). In generating items for a scale, some issues should be considered (Popham 2006). These issues are:

Each item should express only one idea. For example “I like mathematics as a subject and not” “I like mathematics and English Language as school subjects”

Employ the use of both positive and negative statements. This means stating sentences or statements favorably and unfavorably, it could also be in the form of reversing the direction of option given. When this is done, respondents will tarry a while to read through the item before making a conclusion or deciding to choose an option.

Write to the understanding of your respondents to avoid ambiguity. The grammar and sentence structure should be at a level that will be easy for the respondents to understand. Ambiguity should be totally avoided in the drawing of all the items.

The use of sensitive wordings especially those that may bias respondents should be avoided.

Finally, negative wordings of items have to be avoided. In the same manner, the use of double negatives should also be avoided in the items as it is capable of confusing respondents.

Step 4: page layout

Generally, the instructions which will guide the respondents in making correct submissions are located at the top of the page. This is often accompanied by other request for the respondent’s bio-data information. At the top of the first page, clear explanation should be given on the rating format (i.e immediately after the instruction) before the items. The most common rating scale is the type 1-5 type response format commonly used in many research instruments. Where this is used, respondents should understand what 1, 2, 3, 4 & 5 stand for. Not only this, if the questionnaire extends beyond a page, the options (1-5) for rating should be repeated on each page. The use of ambiguous scoring scale should be avoided so as to obtain reliable information and at the same time reduce non- return rate. The instruction and rating options on a scale could be like the one presented for the Work Locus of Control Scale (WLCS); “the following statements concern your beliefs about jobs in general. They do not refer to your present job. Please use the 1 - 6 points scale to rate the extent to which you agree with each statement. 1=Disagree very much, 2=Disagree moderately, 3=Disagree lightly, 4= Agree lightly, 5=Agree moderately and 6=Agree very much”

Step 5: administer the scale

The researcher should submit the scale to expert to review before administering the scale. This is done in order to reduce ambiguities in the use of words, grammar or sentences. The review also helps in identifying structural defects that may arise on format and conceptualization. The scale could be submitted to a number of experts for their professional judgment. Agreement on the issue identified as defects should be sought before a final settlement is reached. The expert or professional judgment though not subjected to any statistical analysis is usually very helpful in having a scale which is capable of capturing the psychological traits of interest.

Sample size

Cohen, (2000) discussed the issue of sample size. He agreed that a large person to item ratio is desirable for robust analysis. When the sample size is small, parametric statistics analyses may not be significant and where occur any significance in analyses, any change in few scores may drastically affect the measure of central tendency. To this end, Tinsel & Tinsel in Cohen (2000), suggested about 5-10 respondents per item up to a maximum of 300 respondents. Comrey (1988) suggested a sample size of about 200 saying that "it is usually acceptable since most scales have no more than 400 items". In the same vein, De vellis (1991) submitted that 400 respondents might be acceptable for the factor analyses of a 90- item scale. It therefore means that authors and researchers are in unison on the issue of sample size (which should be large) for reliable data and robust analysis.

Step 6: check the data

In checking the data investigator is expected to inspect the responses in order to defect the initial respondents' errors on the scale. These will include extreme scores, unusual responses, improperly marked items skipping or omission of items, evidence of confusion and scores outside the range of what is expected. When these are detected, the researcher will have to decide whether to include them or remove them from the data.

This again underscores the importance of large sample size which will make a few 'unusual cases' have little or no effect on the nature of the result. Part of this check is the running of data frequencies (using the statistical products and service solution (SPSS) to check the distribution of each item. This will reveal the skewness and the inter- correlation (Ott and Longnecker, 2001) of the items. Skewed items are likely to be those that are poorly worded and do not discriminate well. The items that have the highest index of inter- correlations will be the ones to be selected and those with poor inter – correlations are to be dropped.

Step7a: coefficient – alpha

The coefficient alpha was put forward by Cronbach (1951) as a measure of internal consistency of items on a scale. It explains the extent to which the items on a scale measure the same underlying property (i.e homogeneity of items). The value of coefficient alpha ranges from 0 to 1, this could assume a positive status for items that are positively correlated and negative when the items do not positively correlate with each other. The coefficient alpha also reveals the proportion of variance it shared with other items when squared.

Olanrewaju (2019) recommended that an alpha of less than 0.60 is unacceptable; 0.60-0.65 undesirable; 0,64-0.70 minimally acceptable;0.71-0.80 respectable; 0.80-0.90 very good; and above 0.90 excellent. Where the alpha coefficient for an item is low, it means the item share little in common with the others on the scale, such item should be discarded and new items generated as replacement. When items with lower correlations are removed from the scale, the homogeneity of the items on the scale is increased, the reliability of the scale is improved and greater confidence is gained in the stability of the measure.

Step7b: factor analysis

After generating item, administering them and determining the alpha coefficient to a sample of the population of interest is factor analysis. The researcher has considered reverse- scored items, the number of items adequate, scaling of the items and adequate sample size. The next is to construct the scale or measure of the construct, through a process of reduction of the number of items and the refinement of the construct.

The most common technique for doing this is factor analysis (Ford, MacCallum & Tait, 1989). Factor analysis addresses the problem of analyzing the structure of the interrelationship (correlations) among a large number of variables (e.g test scores, test items questionnaire responses) by defining a set of underlying dimensions known as factors. In other words, factor analysis is a name given to a group of statistical techniques that can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors).

Factor analysis according to Olanrewaju (2019) is a data reduction technique which assumes that the observed variables are linear combination of some convert, underlying, unobservable or hypothetical factors, and it is used in determining the nature of the underlying /unobservable patterns among a large number of variables. The method is statistical technique which is widely employed by researchers trying to develop measurement instruments or scales. Factor analysis is one of the methods of addressing the question of validity of instruments.

According to Chen et al (2000), researchers rarely find the desired instrument to employ in collecting data for their studies even after a 'significant perseverance in literature'. When the right instrument is obtained from literature, there is a lack of information concerning its psychometric properties which could warrant its use.

The approach involves condensing the information contained in a number of original variables into a smaller set of dimensions (factors) with a minimum loss of information. Factor analysis is different from the dependence techniques e.g multiple regression, in which one or more are explicitly considered the criterion or dependent variables and all others are the predictor or independent variables. It is an interdependence technique in which all variables are simultaneously considered, each related to all others.

To reduce a large number of variables to a smaller number of factors for modeling purposes, where the large number of variables precludes modeling all the measures individually. As such, factor analysis is integrated in Structural Equation Modeling (SEM), helping create the latent variables modeled by SEM. However, factor analysis can be and is often used on a stand-alone basis for similar purposes.

To select a subset of variables from a larger set based on which original variables have the highest correlations with the principal component factors.

To create a set of factors to be treated as uncorrelated variables as one approach to handling multi-collinearity in such procedures as multiple regression.

To validate a scale or index by demonstrating that its constituent items load on the same factor, and to drop proposed scale items which cross- load on more than one factor. To establish that multiple tests measure the same factor, thereby giving justification for administering fewer tests. Factor analysis can be used for exploratory or confirmatory purposes.

Exploratory Factor Analysis

This seeks to uncover the underlying structure of a relatively large set of variables. It is commonly used by researcher when developing a scale and serves to identify a set of latent constructs underlying a battery to measured variables. The researcher's a priori assumption is that any indicator may be associated with any factor. This is the most common form of factor analysis and one which will be covered in this note.

Confirmatory Factor Analysis

This seeks to determine if the number of factors and the loading of measured (indicators) variables on them conform to what is expected on the basis of pre- established theory. The objective of confirmatory factor analysis is to test whether the data fit a hypothesized measurement model. This hypothesized model is based on theory and /or previous analytic research CFA was first developed by Joreskog and has built upon and replaced older methods of analyzing construct validity as described in Campbell & Fiske (1959).

Validation of Rating Scale

Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the purposed use. According to Kolawole (2010), validity has to do with the question of what test scores measure and what they will predict. A score is valid for predicting anything with which it correlates, where "anything" does not include the score itself, for a self-prediction has to do with reliability. When two tests show an inter correlation greater than zero, the "what" that one of them measure is identical, at least in part, with the "what" that the other measure. Adebule and Oluwatayo, (2011) asserted that Validity refers to the degree in which a test or other measuring device is truly measuring what is intended to measure. The test question 1+1 = ……is certainly a valid basic addition question because it is truly measuring a student's ability to perform basic addition. On a test designed to measure knowledge of History, this question becomes completely invalid. The ability to add two single digits has nothing to do with history.

According to Adebule and Oluwatayo (2011), Test validity is the extent to which inferences, conclusion and decision made on the basis of test scores are appropriate and meaningful for many constructs or variables that are artificial or difficult to measure, the concept validity becomes more

complex. Validity is often assessed along with reliability defined as the extent to high a measurement gives consistent results. An early definition of test validity identified it with the degree of correlation between the test and a criterion. Most of us agree that "1 + 1 = _____" would represent basic addition, but does this question also represent the construct of intelligence? Other constructs include motivation, depression, anger, and practically any human emotion or trait. If we have a difficult time defining the construct, we are going to have an even more difficult time measuring it. Construct validity is the term given to a test that measures a construct accurately and there are different types of construct validity that we should be concerned with, among which are concurrent validity and predictive validity. While we speak of the validity of a test or instrument, validity is not a property of the test itself. Instead, validity is the extents to which the interpretations of the results of a test are warranted, which depend on the test's intended use (i.e, measurement of the underlying construct). Relatedly (Whiston, 2005) views validity as the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests.

Similarly, (Kaplan & Saccuzzo, 2005) view validity as the evidence for interferences made about a test score. Furthermore (McBurney & White, 2007) view validity as an indication of accuracy in terms of the extent to which a research conclusion corresponds with reality. The foregoing suggests that validity hinges on the extent to which meaningful and appropriate inferences or decision are made on the basis of scores derived from the instrument used in research. Hypothetical constructs cannot be measured directly and can only be inferred from observations of specified behaviours or phenomena that are thought to be indicators of the presence of the construct.

Measurement of a construct requires that the conceptual definition be translated into an operational definition. An operational definition of a construct links the conceptual or theoretical definition to more concrete indicators that have numbers applied to signify the "amount" of the construct. The ability to define and quantify a construct is the core of measurement.

Validity evidence is built over time, with validations occurring in a variety of populations. Comprehensive reviews of related literature on measurement approaches are therefore critical in guiding the selection of measures and measurement instruments. There are four types of validity that are of much importance and they are: face, content, construct and criterion related validity.

Face validity

This refers to the extent to which a measuring instrument appears valid on the surface (Jackson, 2010). It can also be referred to as researcher's suggestive assessment of the presentation and relevance of measuring instruments as to whether the items in the instruments appear to be relevant, reasonable, unambiguous and clear. Face validity is determined by a cursory review of the items by untrained judges.

Several authors have commented on the status of the face validity in research. Most of these authors believe that face validity is not truly an indicator of validity and hence should not be considered as one (Kapan & Saccuzzo, 2005, Whiston 2005). The argument is that face validity does not offer

evidence to support conclusions drawn from the test scores and that an instrument may look valid without measuring what it is intended to measure.

However, according to Anastasi and Urbina (2007) is a desirable feature of test's noting that if tests originally design for children and developed with a classroom setting are extended for adult's use such tests are likely to meet with resistance and criticism because of their lack of face validity. In other word, if the content of a measuring instrument appears irrelevant, inappropriate, silly or childish, there is every likely hood that the result obtained might provide false information and misleading decision. Practically, the quantitative assessment of face validity can be achieved by achieved by having experts in the field of study or psychometrically on unsophisticated interested persons to rate the suitability of the measuring instrument for it intended use.

The criteria for assessment are as follow: the clarity and unambiguity of items, appropriateness of difficulty level for the respondents, correct spelling of difficult words, spacing of items between lines and legibility of print out, attractiveness of paper used.

Construct validity

This type of validity refers to the extent to which a test captures a specific theoretical construct or trait, and it overlaps with some of the other aspects of validity. Evaluation of construct validity requires examining the relationship of the measure being evaluated with variables known to be related or theoretically related to the construct measure by the instrument. All evidence of validity, including content and criterion- related validity contributes to the evidence of construct validity.

Predictive Validity:

In order for a test to be a valid screening device for some future behavior, it must have predictive validity. The SAT is used by college screening committees as one way to predict college grades. The GMAT is used to predict success in business school. And the LSAT is used as a means to predict law school performance. The main concern with these and many other predictive measures is predictive validity because without it, they would be worthless. We determine predictive validity by computing a correlational coefficient comparing SAT scores, for example, and college grades. If they are directly related, then we can make a prediction regarding college grades based on SAT score. We can show that students who score high on the SAT tend to receive high grades in college.

Content validity

This is the extent a measuring instrument covers a representative sample of the domain of behaviors to be measured (Jackson, 2010). The validity of test content is determined and evaluated by experts in the field. Content validity is a non-statistical type of validity that involves the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured (Anastasi & Urbina, 2007). Content validity evidence involves the degree to which the test matches a content domain associated with the construct. i.e a test of the ability to add two numbers should include a range of combinations of digits. A test with only one digit number or

only even numbers would not have good coverage of the content domain. A test has content validity built into it by careful selection of which items to include (Anastasi & Urbina, 2007). Items are chosen so that they comply either with the test specification which is drawn up through a thorough examination of the subject domain.

CONCLUSION

A Rating scale is a technique through which the observer or rater categorizes the objects, events or persons on a continuum represented by a series of continuous numerals. It can be used to assess the individual once only or on repeated occasions to show developments over time. Among other arears where rating scale could be of help are; for are writing reports to parents, filling out admission blanks for colleges, finding out student needs, making recommendations to employers, supplementing other sources of understanding about the child and stimulating effect upon the individuals who are rated.

REFERENCES

Adebule S.O. (2002): *Development and Validation of Anxiety Rating Scale in Mathematics for Nigerian Secondary Schools*. An Unpublished PhD Thesis. University of Ado Ekiti, Ado Ekiti, Nigeria

Alonge M.F (2004): *Measurement and Evaluation in Education and Psychology.* Second Edition Published and printed by Adebayo printing (Nig) Ltd. Ado – Ekiti Nigeria Bandele, (1994

Anastasi, A. & Urbina, S. (2007). *Psychological testing 2nd ed. Pearson.* NJ: prentice hall

Campbell, S. L., Thompson, C. L., Patriarca, L. A., Luterbach, K. J., Lys, D. B., & Covington, V. M. (2013). The predictive validity of measures of teacher candidate programs and performance: Toward an evidence-based approach to teacher preparation. *Journal of Teacher Education (5), 439-453.* doi: 10.1177/0022487113496431

Chen, G., Gully, S. M., Whiteman, J., & Kilcullen, R, N. (2000). Examination of relationships among trait-like individual differences, state-like individual differences, and learning performance. *Journal of Applied Psychology, 85(6), 835-847*

Cohen, L., Manion, L. & Morrison, K. (2000). *Research Method in Education.* London: Routledge Falmer

Cronbach L.J. (1971): *Test Validation* in R.L Thorndike (Ed), Educational Measurement (2nd ed). Washington, DC: American Council on Education

Cronbach L.J. (1971): *Test Validation* in R.L Thorndike (Ed), Educational Measurement (2nd ed). Washington, DC: American Council on Education

De Ros-Vossles, D., & Fowler-Haughey, S. (Sept. 2007). *Why children's dispositions should matter to all teachers*. Beyond the Journal-Young Children on the Web. Washington DC: National Association for the Education of Young Children

Ford, J.K., MacCallum, R.C., & Tait, M. (1989). *The applications of exploratory factor analysis in applied psychology*: A critical review and analysis. Personnel psychology, 39, 291-314 Greenwich, CT: Information Age. Harcourt, Nigeria.

Hopkins, K.D. (1998). *Measurement and Evaluation in Education and Psychology*. Boston: Allynand Bacon

Jackson, (2010). Validity and Reliability of coaching competency in team sport. *Journal of Physical Education and sport. 16(2):493-499*

Kapan, R.M & Saccuzzo, D.P (2005). *Psychological Testing Principles, applications and issues.* Belmont, CA: Thomson Wadsworth.

Kaplan, R. M., & Saccuzzo, D. P. (2005) *Psychological Testing 1, 2, 3*

Kolawole, E.B. (2005): *Measurement and Assessment in Education*. Published by Bolabay Publications. Ikeja, Lagos, Nigeria

Kolawole, E.B. (2010). *Principles of Test Construction and Administration*. Published by Bolabay Publications. Ikeja, Lagos, Nigeria

Likert, R. (1932): *A technique for the measurement of attitudes*. Archives of Psychology 22(40), 1-55

McBurney, D. H. & White, T. L. (2007). *Research methods 7th ed.* Thomson Wadsworth, 169

Olanrewaju, B. O. (2014). Assessment of the psychometric properties of Mathematics Achievement Test formats. Unpublished M.Ed Thesis, Ekiti State University, Ado Ekiti

Olanrewaju, B.O. (2019): Construction and Validation of Disposition Scale of University Lecturers towards Teaching Profession in Southwest, Nigeria. Unpublished Ph.D. Thesis. Ekiti State University, Ado-Ekiti Cohen, L., Manion, L. & Morrison, K. (2007). *Research Method in Education.* London: Routledge Falmer

Omirin M.S (1999): *Construction and Validation of Science oriented Attitudinal Scale for Nigerian Schools*. An unpublished PhD thesis. University of Ado Ekiti, Ado Ekiti Nigeria

Ott, RL. & Longnecker, M. (2001). *An Introduction to statistical methods and data analysis* (5th. Ed) Durbury: Thamson Learning

Robert D. (1996): Design and Construction of Psychological Measures. Microsoft Internet Explorer

Whiston, S. C. (2005). Principles and applications of assessment in counseling 2nd ed. Thomson Brooks/Cole. 43-74.