International Journal of Studies in Psychology
E-ISSN: 2710-2327, P-ISSN: 2710-2319
Vol 3, No. 2, pp 77-81. https://doi.org/10.38140/ijspsy.v3i2.941
GAERPSY Publishing, 2023

History Article
Submitted 13 September 2023
Revised 30 September 2023
Accepted 10 October 2023
Published 15 November 2023

Dimensionality of chemistry teachers' effectiveness scale (CTES) in secondary schools in Osun state, Nigeria

Chidubem Deborah Adamu
Obafemi Awolowo University, Ile-Ife, Nigeria
Primary author: adamuchidubemdeborah@gmail.com

***Abstract*—** The study assessed the dimensionality of the Chemistry Teachers' Effectiveness Scale (CTES) in Osun State, Nigeria secondary schools. Also, the study determined the extent to which CTES satisfies the unidimensionality assumption of the Item Response Theory model. It determined the extent to which the observed unidimensionality of CTES was confirmed when the scalability of the individual items and the overall scale was assessed. The study employed a survey research design. Thirty-five (35) Chemistry teachers who were rated by their Heads of Departments and Chemistry students made up the sample for the study. A multistage sampling procedure was employed for choosing the sample in two phases-: validation (conducted in Oyo State), and pilot testing (conducted in Osun State). A self-developed research instrument titled "Chemistry Teachers Effectiveness Scale (CTES)" was used for data collection. Items of the CTES were rated on a four-point Likert-type scale described under 1 = very poor, 2 = poor, 3 = moderate, and 4 = good. The Chemistry teachers' effectiveness scale started with the initial version of 206 items, and was reduced to 96 items after validation. The 96-item second version of CTES was reduced to 62 items after pilot testing and reliability analysis (giving Cronbach Alpha coefficient of 0.92). A thorough and more robust statistical analysis was conducted on the 62-item third version of CTES. Mokken Scaling Analysis (MSA) was used to analyse the data via the Mokken package. Results showed that all the items of CTES have scalability coefficients within the 0.20 and 0.39 range and violated the Item Response Theory (IRT) unidimensionality assumption. The study concluded that items of the CTES are multi-dimensional.

Keywords: Chemistry Teachers Effectiveness Scale, dimensionality, Item Response Theory, Multidimensionality, Unidimensionality

To cite this article (APA): Adamu, C. D. (2023). Dimensionality of chemistry teachers' effectiveness scale (CTES) in secondary schools in Osun state, Nigeria. *International Journal of Studies in Psychology*, *3*(2), 77-81. https://doi.org/10.38140/ijspsy.v3i2.941.

I. INTRODUCTION

IT is a crucial issue in research for every study to employ an instrument. A research instrument, therefore, can be a scale that the researcher designs, adapts, or adopts to collect data from the study participants. A scale is an analytical tool that assembles various questions or statements merged into an overall score and aimed at unmasking abstract ideas that cannot be ordinarily seen by merely looking at them (Devellis, 2012). Any easily understood idea that is not obvious but can be known or perceived from empirical findings can be referred to as a hypothetical construct (Fon & Cahill, 2001, cited in Neil, Smelser & Baltes, 2001). In psychology, many guessed inward activities are metaphysical constructs expected to validate certain observable actions. In other words, a theoretical construct can also be referred to as a hypothesised construct or phenomenon which is not directly observable. It is based upon certain characteristics (visible) that make up a phenomenon. For instance, if a candidate is seen sweating in an examination, it can be fairly deduced that they are having anxiety. This anxiety from the examination is a form of social construct. Even though it cannot be observed directly, data on it can be collected using paper and pencil measurement scales such as questionnaires. Other examples of theoretical or psychological constructs are love, emotion, hate, anxiety, consciousness, effectiveness, intelligence, personality, and attitude.

If and when researchers find out concepts that are not easily measured directly, factor analysis allows them to collapse a large number of variables into a few interpretable underlying factors such as behaviours shown by teachers in the classroom. These behaviours determine their effectiveness. As a statistical tool used for data reduction or structure detection, factor analysis seeks unobserved variables that are reflected in the manifest variables.

Negative and positive behaviours exhibited by teachers affect their effectiveness in the classroom (Stronge, 2018). Some of these behaviours [teachers' personality, classroom management and organisation, organising and orienting for instruction, implementation of teaching, monitoring students' progress and potential; and professional development] (Stronge, 2018) are not intended but overtly reflect themselves in a teacher. They make up the teacher qualities such as their personality, ways of managing and organising the classroom, organising and planning for lesson, actual execution of the lesson; follow-up on students' progress and potentials; and self-development in the teaching profession.

Teacher effectiveness therefore, is the overall qualities teachers possess that enable them to influence students learning and their achievement in standardised tests (Badri & Al-Khaili, 2014). In this study, supervisors' rating and Chemistry students' reports were adopted in measuring the dimensionality of the items of CTES. This is to corroborate Ajeigbe and Afolabi (2014) that in psychometrics, test experts are expected to generate a set of quality test items forming an

instrument that must measure just one thing in common. In the Rasch model, the dimensionality of a research instrument can be either unidimensional (a research instrument measuring only one item at a time) or multi-dimensional (when two or more items in a test are measured simultaneously).

Unidimensionality and local independence are the two most critical and basic assumptions of the Rasch model. Local independence assumes that the response given by an examinee in a test or measuring instrument must not be influenced by his or her response to another item in the same test (that is, all responses given to all items in a test are independent of one another). Unidimensionality assumes an item should measure one attribute at a time (Babatimehin, Adamu & Adeoye, 2021). The assumption provides the basis for most mathematical measurement models. In making psychological sense when relating variables, ordering persons on some attribute, forming groups based on some variable; or making comments about individual differences, the variable must be unidimensional. This is to say that the various items in an instrument measure the same ability, achievement, attitude, or other psychological variables (Guler, Uyanik, & Teker, 2014). Unidimensionality and local independence are determined through fit statistics (Sick, 2010 as cited in Ajeigbe & Afolabi, 2014). Fit statistics report the degree to which the pattern of observed responses and the modeled outcomes are evaluated regarding item fit and person fit to the Rasch model. In the Item Response Theory (IRT) model of the Rasch model, a strong relationship exists between unidimensionality and local independence assumptions. For instance, unidimensionality is encountered in a test when individual items measure one construct at a time. At the same time, in local independence, the response to each item in a test is not affected by the response to another item. It implies that attention is on individual items in unidimensionality and local independence assumptions. If and when unidimensionality is absent in a test, multidimensionality is suspected. The focus of this study is the dimensionality of CTES.

Most factor analysis models assume that an individual's response to a test item is influenced by only one dimension. This is such that cross-loadings are usually deemed as item vagueness and lack of objectivity. On the contrary, multi-dimensional within-item models comprise items relating to more than one dimension at a time and require a more complex loading structure to model the relationship between the latent traits under investigation. These models allow for cross-loading items, do not treat the dimensions as separate scales, and provide a more robust insight into convergent and discriminant validity. As opposed to between-item multi-dimensional models, within-item multi-dimensional frameworks employ methods that take full advantage of the information in the data, which is against depending on scarce information. These models are, therefore, called "full information" models (Bock, Gibbons & Muraki, 1988) since they are based on individual's pattern of response rather than on the correlational structure of the multivariate latent response distribution (Wirth & Edwards, 2007). Mokken (1971) as cited in Sijtsma and Molenaar (2002); and Van-Abswoude, Van-Der, and Sijtsma (2004) recommended that sub-factors are interpreted based on the items loading on them (while sub-factors loading three or more items are retained in a scale, factors loading less than three items are deleted from a scale). Multidimensional Item Response Theory (MIRT) is an extension of the unidimensional IRT model (Reckase, 2009). The extension of IRT models to within-item models is called MIRT.

In this study, the Chemistry teachers' effectiveness scale is a Likert-type scale in which the items are polytomously scored. As a result, the Multi-dimensional Graded Response Model (MGRM) was employed for the study (Carlson, 1987). The multi-dimensional graded response model generalises traditional polytomous IRT models and considers item difficulty and discrimination. However, MGRM assumes that selecting a response category requires several steps, and reaching step k requires acceptance of k - 1.

Statistically, determining the dimension of a scale is very crucial and should be a rigorous procedure. In the IRT, two dimensions of a test exist: unidimensional and multi-dimensional. Although, researchers usually endeavour that each item in a test measures a single construct at a time. In most cases, it is observed that virtually, all the item scores derived from these measures fail to meet the strict goodness-of-fit criteria of each measuring one trait required by the single-factor analysis model (Garrido, Gonzalez, Seva, & Piera, 2019). In this cases, it is predicted to fit multiple correlated factor analysis solutions to the data and propose the resulting solutions (which fit the model most) as the most appropriate structures for the measures under scrutiny (Ferrando & Lorenzo-Seva, 2018a, 2018b; Furnham, 1990; Reise, Bonifay, & Haviland, 2013; Reise, Cook, & Moore, 2015). However, most instruments designed to measure a single construct yield data that is compatible with a solution in which there is a strong, dominant factor running through all the test items (Floyd & Widaman, 1995; Reise, Bonifay & Haviland, 2013; Reise, Cook & Moore, 2015). So, when dimensionality is being judged, the emphasis should not be (or not be) on the goodness-of-fit and factor structure of the solution but rather on the properties of the score estimates derived from this solution (Ferrando & Lorenzo-Seva, 2018b, Ferrando & Navarro-Gonzalez, 2018). According to Garrido, Gonzalez, Seva, and Piera (2019), there are consequences for adopting a wrong decision on the dimension of a measuring instrument. For instance, if item scores are essentially unidimensional but are treated as multi-dimensional, the main potential consequences are: lack of clarity in the interpretation and unnecessary theoretical complexities; weak; nonreplicable factors of little substantive interest; and (as a consequence) weakened factor score estimates that do not allow accurate individual measurements to be made. On the other hand, treating multi-dimensional scores as unidimensional is expected to lead to biased item parameter estimates, loss of information, and factor estimates that cannot be univocally interpreted because they reflect the impact of multiple sources of variance (Ferrando & Navarro-Gonzalez, 2018; Reise, Bonifay & Haviland, 2013; & Reise, Cook & Moore, 2015). Margono (2015) claims that most psychological tests and scales are multi-dimensional measurements rather than unidimensional in cognitive and affective measures.

Despite unidimensionality's importance, there is no accepted and effective index for its set of items. Lord (1980) affirmed that such an index is of great importance, while Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978) disagreed that assessing the assumption of unidimensionality outweighs other goodness-of-fit tests under a latent trait model. From the preceding, multiple methods have been suggested in testing unidimensionality in developed countries. They are test essential dimensionality (Nandakumar & stout, 1993; Stout, 1987), Bootstrap modified parallel analysis test, full information item factor analysis (Bock, Gibbons & Muraki, 1988), Exploratory Factor Analysis (EFA) of tetrachoric correlations (Knol & Berger, 1991), confirmatory factor analysis of tetrachoric correlations with robust weighted least squares estimation (Muthen, 1993), non-linear factor analysis (McDonald, 1967; 1962) and many others. Similarly, Hattie, Krakowski, Rogers, and Swaminathan (1996) conducted a simulation study to evaluate the dependability of Stout's unidimensionality index as used in his DIMTEST procedure. The results showed that DIMTEST dependably provides indications of unidimensionality, is reasonably robust, and allows for practical distinction between and among many dimensions.

Traditional factor analysis used in SPSS is mostly employed in testing the unidimensionality of tests in Nigeria (Matibemu, Oguoma & Essien, 2019). For instance, Ojerinde (2012) consented that the following are the methods for estimating unidimensionality in a test: Eigenvalue test, congruence test, Structural Equation Modelling (SEM) test, factor analysis, Cronbach analysis test, Vector frequency test and Confirmatory Factor Analysis (CFA), commonality test, part/whole test, factor loading test, random baseline test, and biserial test. When using the Cronbach alpha, the coefficient alpha (α) must be greater than or equal to 0.70 (Ojerinde, 2013). Nevertheless, if a value greater than

0.70 is obtained, the items in the test are unidimensional but multi-dimensional if on the contrary. Ojerinde (2013) assessed the unidimensionality of UTME mathematics pretest, which was dichotomously scored using Eigenvalues analysis via scree plot and Kuder-Richardson 20 formula. The author adjudged the test unidimensional because the KR20 was higher than 0.7. At the same time, the scree plot of the Eigenvalue showed that the first Eigenvalue was larger compared to the second factor, and the Eigenvalue of the remaining factors were all about the same.

Similarly, Ajeigbe and Afolabi (2014) assessed unidimensionality and occurrence of Differential Item Functioning (DIF) in Mathematics and English Language items of the Osun State Qualifying Examination (OSQE) using secondary data. In the study, results showed that OSQE Mathematics ($-0.094 \leq r \leq 0.236$) and English Language ($-0.095 \leq r \leq 0.228$) were unidimensional; there was occurrence of DIF items in both Mathematics and English Language multiple-choice items of the OSQE for 2008. The authors concluded that the examination contained many items that exhibited DIF and, therefore, requires adequate item quality improvement to justify its use as the inclusion or exclusion criterion of state candidates in the West African Examination Council. Nevertheless, Metibemu (2016) assessed the unidimensionality of a 100 item Mathematics achievement test using EFA performed in Statistical Packages for Social Sciences (SPSS). The first and the second Eigenvalues of the components were extracted using factor analysis in the study. The study also concluded that the test was unidimensional and satisfied the unidimensionality assumption of the IRT as the ratio of the first Eigenvalue to the second Eigenvalue was greater than one. More so, Awopeju and Afolabi (2016) pronounced the 2008 NECO Mathematics multiple-choice test unidimensional. This was because the ratio of the first to the second Eigenvalues of the tests extracted using factor analysis of SPSS version 20 was 3 to 1. None of the studies reviewed in the current study assessed the dimensionality of Chemistry teachers' effectiveness scale in secondary schools in Osun state, Nigeria. Assessing the dimensionality of a test or scale involves rigorous statistical procedure and expertise.

It has become a norm among most test developers to adjudge the items in a unidimensional test without carrying out a more rigorous and statistical analysis. Almost entirely, all the item scores obtained from unidimensionality measures are proven inadequate to meet the strict goodness-of-fit benchmark and factor structure of an individual item measuring a single construct. Notwithstanding, in assessing the dimensionality of a scale, emphasis should not or be on the goodness-of-fit and factor structure. Pronouncing a test unidimensional entails that it must strongly satisfy the unidimensionality assumption of the IRT; if violated, then multidimensionality is suspected. However, different IRT models have evolved and been used to test data but are usually adjudged unidimensional among most psychometricians. It is consequential if and when item scores are necessarily unidimensional but are considered multi-dimensional. In a real sense, one would ask if all the items in a test are measuring a single construct at a time. The answer may be no. If that is the case, then it can be assumed that some procedural errors may be accounted for in the statistical analysis method used in measuring such tests or scales. Literature has proven that the most correct statistical analysis must be concluded to determine tests' dimensionality. This study, therefore, questions the place of the multi-dimensional counterpart of the unidimensionality of test items.

II. OBJECTIVES OF THE STUDY

The study assess the dimensionality of CTES in secondary schools in Osun State, Nigeria. This was to adjudge the scale's dimensionality in the study area. Specifically, the objectives of the study were to:

1. determine the extent to which CTES satisfies the unidimensionality assumption of the IRT model; and

2. determine the extent to which the observed unidimensionality of CTES was confirmed when the scalability of the individual items and the overall scale were assessed.

The following research questions were raised from the above-stated objectives.

1. To what extent does CTES satisfy the unidimensionality assumption of the IRT?

2. To what extent is the observed unidimensionality of CTES confirmed when the scalability of the individual items and the overall scale were assessed?

III. METHODS

The study adopted the survey research design. The population for the study consisted of all Chemistry teachers in all the federal-, State-, and privately owned secondary schools in Osun and Oyo States, Nigeria. The sample for the study comprised 35 Chemistry teachers who were rated by their Heads of Departments and Chemistry students. Multistage sampling technique was used to select the sample in two phases: validation (carried out in Oyo State) and pilot testing (done in Osun State). In phase one, a sample for establishing the face and content validity of the items of CTES was established in Oyo State. Here, four experts reviewed an initial 206 items of CTES. At this point, while 88 items were found not to reflect the true purpose of the scale, 22 items were double-barrelled (totalling 110 items). These 110 items were deleted, leaving the initial scale with 96 items. In phase two, a purposive sampling technique was employed in the selection of all the Chemistry teachers in the three Federal Government Colleges (totalling 13 Chemistry teachers), state-owned secondary schools (rounding up seven), and privately-owned secondary schools (summing up three) in Osun State. This implies that the total sample used in phase two was 23 Chemistry teachers. In phase two, pilot testing was carried out on the 23 Chemistry teachers. Also, the data collected in phase two was subjected to Exploratory Factor Analysis (EFA) to select items. At this stage, thirty-four items that did not meet one of the criteria of factor loadings of 0.5 and above were eliminated. Therefore, 96 items of the CTES were reduced to 62. The scale's reliability was determined with Cronbach Alpha and returned a high-reliability coefficient of $r = 0.93$. This value indicates that CTES has a good reliability coefficient. The research instrument used for data collection was self-developed and titled "Chemistry Teachers Effectiveness Scale (CTES)." Items of the CTES were rated on a four-point Likert-type scale described under 1 = very poor, 2 = poor, 3 = moderate, and 4 = good. The 62-item CTES was subjected to robust statistical Mokken Scaling Analysis, a non-parametric IRT model.

IV. RESULTS AND DISCUSSION

Research Question One: To what extent does CTES satisfy the unidimensionality assumption of the IRT model? To answer this research question, the responses of Osun State Chemistry teachers to the CTES were subjected to MSA. The analysis assesses the measurability of a scale. To assess the scalability of the scale, three measures were used. They are: Loevinger's scalability coefficients for item-pair (Hij), Loevinger's scalability coefficients for the item (Hi), and Loevinger's coefficients for scale (Hs). The MSA of the data was conducted with the Mokken package. First, the possibility of all the items on the scale measuring a single construct was assessed. The items on a scale measure a single construct when the item's scalability coefficient of each pair (Hij) is positive. If unidimensionality is not tenable by the scale, the number of dimensions underlying the scale and the scalability of each dimension or construct are evaluated. The rules of thumb for adjudging the scalability of the scale are as follows: a scale is weak if $0.3 \leq H < 0.4$, moderate if $0.4 \leq H < 0.5$, and strong if $H \geq 0.5$ (Lee, Fu, Liu & Hung, 2017). The result is presented in Table 1.

Table 1 shows the assessment of the unidimensionality of CTES. The Table showed that the scalability coefficient of each pair (Hij) of the items on the scale returned a positive value. The result showed that the Chemistry teachers' effectiveness scale measured only one construct. To

confirm the observed unidimensionality of the CTES, the scalability of the individual items and the overall scale were assessed. The result is presented in Table 2.

Research Question Two: To what extent is the observed unidimensionality of CTES confirmed when the scalability of the individual items and the overall scale were assessed?

Table 2 shows the ability of the 62-item CTES to form a single-factor scale. The Table showed that all the items of CTES have scalability coefficients within the 0.20 and 0.39 range. This shows that the items were weak measures of a single-factor CTES. Similarly, the overall scalability coefficients of the CTES (H = 0.31) were below the 0.40 minimum benchmark (Lee, Fu, Liu & Hung, 2017) for considering a scale a moderate measure of the constructs it is designed to measure. The results suggest that the single factor CTES is a "weak" measure of Chemistry teachers' effectiveness.

In research question one, it was found that the 62 items of CTES seemed to be unidimensional. The items looked like they were measuring the same trait as all the scalability coefficients of each pair (Hij) of the items on the scale gave back a positive value. This is in line with Garrido, Gonzalez, Seva, and Piera (2019), who posited that there are consequences for adopting a wrong decision on the dimension of a measuring instrument. As a result, items on the scale were subjected to further statistical analysis.

More so, findings in research question two showed the scalability of the individual items and the overall scale. Again, it showed that the 62 items of the CTES were weak measures of the scale, as all the items were less than 0.4. This finding of the study conforms with Lee, Fu, Liu, and Hung (2017) that the rules of thumb for judging the scalability of a scale are as follows: a scale is weak if $0.3 \leq H < 0.4$, moderate if $0.4 \leq H < 0.5$, and strong if $H \geq 0.5$. Hence, the overall scalability coefficient of the CTES was below the recommended value of 0.40 for labeling a scale moderate measure of the traits it purported to measure. This implies that items of CTES are multi-dimensional and not unidimensional.

V. CONCLUSION AND RECOMMANDATIONS

Based on the findings of the study, it was concluded that items of the CTES violated the unidimensionality assumption of the IRT model as each of the validated 62 items of CTES were weak measures of the scale and were less than the acceptable benchmark for adjudging a scale unidimensional. Therefore, items of CTES are adjudged multi-dimensional. However, six scalable factors underlie CTES.

Hence, it was recommended in the study that researchers and psychometricians who are interested in the area of scale development should assess first the dimensionality of the instrument under investigation, and extra care should be taken in labeling a scale unidimensional as a unidimensional instrument may fit a multi-dimensional scale after confirmatory analysis. The researcher acknowledges the research Assistants who made themselves available during the data collection period of the study. Also, the researcher's sincere appreciation goes to all the secondary school Chemistry teachers, their students, and the supervisors who made up the study's participants. Lastly, the researcher appreciates all the Principals and Vice Principals of these secondary schools for granting her consent to conduct the study.

REFERENCES

Ajeigbe, T. O., & Afolabi, E. R. I. (2014). Assessing unidimensionality and differential item functioning in qualifying examination for senior secondary school students in Osun state, Nigeria. *World Journal of Education*, 4(4), 30-37. http://doi.org/10.5430/wje.v4n4p30.

Awopeju, B. P., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory-based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*(28), 263-284. http://doi.org/10.19044/esj.2016.v12n28p263.

Babtimehin, T., Adamu, D. C., & Adeoye, O. P. (2021). Comparison of local item independence of WAEC and NECO dichotomously scored chemistry items in Osun State, Nigeria. *Nigerian Journal of Educational Research and Evaluation, 20,* 194-211.

Badri, H., & Alkhaili, M. (2014). Teacher effectiveness and organizational commitment. *Journal of Academic Ethics, 7*(4), 297-314.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological measurement*, *12*(3), 261-280. https://doi.org/10.1177/014662168801200305.

Carlson, J. E. (1987). *Multi-dimensional item response theory estimation: A computer program*. Iowa City IA: American College testing program. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/ACT_RR87-19.pdf.

Devellis, R. F. (2012). *Scale development: Theory and applications*. Sage Publications. https://doi:10.7334/psicothema.

Ferrando P. J., & Lorenzo-Seva, U. (2018a). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, *78*(5), 762-780. https://doi.org/10.1177/0013164417719308.

Ferrando P. J., & Lorenzo-Seva, U. (2018b). On the added value of multiple factor score estimates essentially unidimensional models. *Educational and Psychological Measurement, 78,* 762- 780. https://doi.org/10.1177/0013164417719308.

Ferrando P. J., & Navarro-Gonzalez, D. (2018a). Assessing the quality and usefulness of factor analytic applications to personality measures: A study with the statistical anxiety scale. *Personality and Individual Differences*, *123,* 81-86. https://doi.org/10.1016/j.paid.2017.11.014.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299. https://doi.org/10.1037/1040-3590.7.3.286.

Foa, E. B., & Cahill, S. P. (2001). Psychological therapies: Emotional processing. In N. J. Smelser & B. Baltes (Eds.). *International Encyclopedia of the Social and Behavioural Sciences* (pp. 12363-12369). Oxford: Elsvier. https://doi.org/10.1016/B0-08-043076-7/01338-3.

Furnham, A. (1990). The development of single trait personality theories. *Personality and Individual Differences*, *11*(9), 923-929. https://doi.org/10.1016/0191-8869(90)90273-T.

Garrido, C. C., Gonzalez, D. N., Seva, U. L., & Piera, P. J. F. (2019). Multi-dimensional or essentially unidimensional? A multi-faceted factor analytic approach for assessing the dimensionality of tests and items. *Psicothema*, *31*(4), 450-457. https://doi.org/10.7334/psicothema2019.153.

Guler, N., Uyanik, K. G., & Teker, G. T. (2013). Comparison of classical test and item response theory in terms of item parameters. *International Journal of Social Sciences Research, 2*(1), 1-6.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent train theory: Models, technical issues, and applications. *Review of Educational Research, 48*(4), 467-510. https://doi.org/10.3102/00346543048004467.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, *20*(1), 1-14. https://doi.org/10.1016/0191-8869(90)90273.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multi-dimensional item response models. *Multivariate Behavioural Research, 26*(3), 457-477. https://doi.org/10.1207/s15327906mbr2603_5.

Lee, C. P., Fu, T. S., Liu, C. Y., & Hung, C. J. (2017). Psychometric evaluation of the Oswerty disability index in patients with chronic low back pain: Factor and Mokken analyses. *Health and Quality Life Outcomes*, *15*(1), 92. https://doi.org/10.1186/s12955-017-0768-8.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hildale, NJ: Lawrence Erlbaum.

Margono, G. (2015). Multi-dimensional reliability of instrument for measuring students' attitude towards statistics by using semantic differential scale. *American Journal of educational Research*, *3*(1), 49-53.

McDonald, R. P. (1962). A general approach to non-linear factor analysis. *Psychometrika, 27*(4), 397-415. https://doi.org/10.1007/BF02289646.

Mcdonald, R. P. (1967). Factor interaction in non-linear factor analysis. *ETS Research Bulletin Series, 1967*(2), i-18. https://doi.org/10.1002/j.2333-8504.1967.tb00990.x.

Metibemu, M. A. (2016). *Comparison of classical test and item response theory in the development and scoring of senior secondary school mathematics tests in Ondo State, Nigeria* (Unpublished doctoral thesis). University of Ibadan, Ibadan, Nigeria.

Metibemu, M. A., Oguoma, C. C., & Essen, C. B. (2019). Ensuring quality in unidimensionality assumption assessment: Evaluating the appropriateness of using traditional factor analysis for multiple-choice test. *Nigerian Journal of Educational research and evaluation*, *18*(1), 206-224.

Mokken, R. J. (1971). *A Theory and Procedure of scale Analysis.* Berlin, Germany: De Gruyter.

Muthen, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.). *Testing structural equation models* (pp. 205-243). Newbury Park, CA: Sage Publications.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*(1), 41-68. https://doi.org/10.3102/10769986018001041.

Ojerinde, D. (2013). Classical test theory (CTT) versus item response theory: An evaluation of the comparability of item analysis results. *A guest lecture presented at the Institute of Education,* University of Ibadan. Retrieved from https://ui.edu.ng/sites/.../prof%20ojerinde's%20lecture%20.pdf.

Ojerinde, O. O. (2012). *General principles of test planning. Educational tests and measurement.* Ile-Ife, Nigeria: Obafemi Awolowo University Press

Reckase, M. (2009). *Statistics for Social and behavioural Sciences: Multi-dimensional Item Response Theory.* Dordrecht: Springer.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129-140. Rtrieved from https://doi.org/10.1080/00223891.2012.725437.

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In Reise S. P. & Revicki D. A. (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13–40). New York: Routledge. https://doi.org/10.4324/9781315736013.

Sick, J. (2010). Rasch measurement in language education part 5: Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 14*(2), 23-29.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory.* Thousand Oaks, CA: Sage.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617. http://doi.org/10.1007/BF02294821.

Stronge, J. H. (2018). *Qualities of effective teachers* (3rd ed.). Alexandria, Virginia: ASCD.

Van-Abswoude, A. A. H., Van-Der, A. L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 4-24. https://doi.org/10.1177/0146621603259277.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58-79. https://doi.org/10.1037/1082-989X.12.1.58.

Table 1: Item-Pair Scalability Coefficient

Item	Hij	Item	Hij	Item	Hij	Item	Hij
1	1.00	16	0.31	31	0.33	46	0.27
2	0.65	17	0.36	32	0.36	47	0.21
3	0.54	18	0.29	33	0.36	48	0.27
4	0.41	19	0.48	34	0.32	49	0.29
5	0.41	20	0.27	35	0.20	50	0.27
6	0.41	21	0.34	36	0.24	51	0.23
7	0.43	22	0.20	37	0.21	52	0.23
8	0.39	23	0.22	38	0.27	53	0.21
9	0.41	24	0.29	39	0.26	54	0.24
10	0.34	25	0.39	40	0.33	55	0.18
11	0.35	26	0.27	41	0.25	56	0.18
12	0.33	27	0.34	42	0.18	57	0.14
13	0.36	28	0.36	43	0.28	58	0.21
14	0.32	29	0.38	44	0.28	59	0.23
15	0.34	30	0.37	45	0.36	60	0.14
						61	0.14
						62	0.20

Table 2: Scalability of the Chemistry Teachers' Effectiveness Scale's Items and the Entire Scale (Source: Author's Analysis, n.d.)

Item	H	Item	H
1	0.31	IT32	0.39
2	0.33	IT33	0.37
3	0.33	IT34	0.35
4	0.27	IT35	0.31
5	0.28	IT36	0.29
6	0.28	IT37	0.31
7	0.35	IT38	0.33
8	0.30	IT39	0.31
9	0.36	IT40	0.30
10	0.31	IT41	0.28
11	0.32	IT42	0.28
12	0.34	IT43	0.29
13	0.31	IT44	0.33
14	0.31	IT45	0.32
15	0.30	IT46	0.31
16	0.28	IT47	0.25
17	0.36	IT48	0.33
18	0.33	IT49	0.34
19	0.34	IT50	0.34
20	0.28	IT51	0.31
21	0.32	IT52	0.29
22	0.25	IT53	0.31
23	0.22	IT54	0.26
24	0.33	IT55	0.26
25	0.36	IT56	0.20
26	0.33	IT57	0.20
27	0.31	IT58	0.25
28	0.35	IT59	0.28
29	0.33	IT60	0.25
30	0.35	IT61	0.21
31	0.36	IT62	0.23
Overall scale	0.31		