



Proof of Unidimensionality in Cognitive Test Instrument for Evaluation Science Learning

Rizki Nor Amelia^{1*}, Septiko Aji¹, Kriswantoro², Hesty Sukmasari³

¹Science Education Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

²Chemistry Education Study Program, Faculty of Teacher Training and Educational Sciences, Universitas Jambi, Indonesia

³Doctoral of Educational Management, Faculty of Educational Sciences and Psychology, Universitas Negeri Yogyakarta, Indonesia

*Correspondence to: rizkinoramelia@mail.unnes.ac.id

Abstract: Dimensional analysis to determine the construct that makes up the instrument as one of the requirements before the analysis is often ignored. Proving dimensions is important because the correct specification of the number of latent dimensions is directly related to construct validity and will determine the appropriate software, approach, and analysis model parameters. Therefore, this research aims to prove the unidimensionality of the Cognitive Test Instrument for Evaluation of Science Learning (CTIESL) through Exploratory Factor Analysis with the Principal Component Analysis Extraction Method. The instrument consisted of 13 essay questions and was responded to by science pre-service teachers who took the Science Learning Evaluation Course (98 people). The results of the analysis prove that the CTIESL has a unidimensional construct based on the Scree Test criteria (detected only one point before the elbow on the scree plot) and the cumulative percent of variance extracted, which shows a satisfactory value (79.232%) so that this instrument is proven only to measure one latent trait, namely the cognitive ability to evaluate the science learning. By fulfilling this unidimensional assumption, the psychometric characteristics of the CTIESL can be explored further with Item Response Theory: 1-Logistic Parameters, 2-Logistic Parameters, or 3-Logistic Parameters, depending on needs.

Keywords: exploratory factor analysis; principal component analysis; unidimensional

Recommended citation: Amelia, R. N., Aji, S., Kriswantoro., & Sukmasari, H. (2025). Proof of Unidimensionality in Cognitive Test Instrument for Evaluation Science Learning. *Journal of Innovation in Educational and Cultural Research*, 6(1), 16-24.

INTRODUCTION

In the teaching and learning process, tests are administered with the expectation that the measurements obtained from the tests will assist in decision-making with minimal risk (Ohiri & Okoye, 2023). Quantitative and qualitative analysis of test items determines their characteristics and improves the instrument (revising or discarding items that do not meet the cut-off) (Ayanwale et al., 2022). Two relevant item analysis approaches are Classical Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory has established estimation hypotheses for over 80 years (Wagbara & Nna-Kue, 2024). This theory is often referred to as true-score theory (Cappelleri et al., 2014) or classical pure score theory (Allen & Yen, 1979), is a conventional quantitative approach that assumes a relationship between observed score (X), true score (T), and random error score (E) in an equation Classical test theory has been widely used in the process of analyzing test items because it can work on small samples so the statistics are simple and the assumptions are not too complicated (Mardapi, 2012).

Traditionally, testee abilities tend to be reported in the form of the number of questions answered correctly. This is a limitation or weakness of CTT, where testees with different response patterns but the same number of correct results will provide the exact ability estimate (Eleje et al., 2018). Apart from that, there is a possibility that a testee with a lower score on a test will have a higher position in the latent trait due to lucky guessing or cheating, which results in parameter invariance (dependence of parameter estimates, for example, the difference power and difficulty level of items on the sample) (Rusch et al., 2017). Therefore, using the number of observed scores is inappropriate for measuring empirical reality (Rusch et al., 2017). In the end, the shortcomings of CTT were successfully overcome by IRT, which allows non-linear relationships, true score estimation that is independent of the sample used, parameter estimation that is invariant to sample values, and facilitates measurement models according to the type of item parameters desired for analysis (Ayanwale et al., 2022). Item Response Theory provides a more useful framework in educational measurement because it supports useful activities such as instrument construction, linking and equating measurements, evaluation of item bias and test bias, and its foundation in item banking and test administration design and flexible (for example, multiple matrix sampling, flexi-level testing, or computerized adaptive testing) (Glas, 2008). There are three hypotheses underlying IRT. First, the unidimensional hypothesis of latent space which assumes that if the

testee's response to all test questions involves n latent traits (or abilities), then these n latent traits constitute an n -dimensional space; second, the local independence hypothesis which refers to the responses of subjects who have the same level of ability to different test questions, are independent of each other, and the subject's responses to one test question are also not influenced by their reactions to other questions; third, the hypothesis about the item characteristic curve (Item Characteristic Curve) which describes a specific graphical representation of the functional relationship between the probability of a correct answer to a test item and the level of ability or trait that corresponds to that item (Hu et al., 2021; Erguven, 2013).

The first and main assumption, namely unidimensionality, turns out to be an assumption that needs to be fulfilled in both IRT (Hagell, 2014; Hu et al., 2021; Toni et al., 2004) and CTT (Hagell, 2014; Mardapi, 2012; Toni et al., 2004). The unidimensionality assumption states that there is only one latent trait or dominant ability that is measured in the test (Bichi & Talib, 2018; Erguven, 2013; Pasquali, 2007; Pasquali, 2003), so it can be concluded that the testee's performance only depends on one factor (Mitee, 2019). The concepts of unidimensional assumption and local independence are similar (when the unidimensional assumption is met, the local independence assumption is also automatically fulfilled) (Erguven, 2013). However, the two are not equivalent because the local independence assumption can still be met without unidimensional data as long as all the influencing aspects of test results can be considered (McBride, 2001). This shows that the unidimensional condition does not mean the items must be positively correlated. An item may be negatively correlated with others and still have one dimension (Bichi & Talib, 2018).

Many IRT-based tests do not meet this assumption; for example, a history test instrument may contain several items about government, technology, arts, culture, etc. Tests like this are not completely unidimensional but must be close enough to use IRT effectively (Livingston, 2020). There are various methods for proving the unidimensional assumption that is not claimed to be superior because each method has its characteristics. According to Hattie (1985), the methods used to evaluate the number of dimensions can be categorized based on (1) Answer Patterns (Green RepB, Consistency, Loevinger's index); (2) Reliability (Alpha, Horst's Alpha, Raju's Alpha, Mean Item Correlation, Mean Correlation, No. Zero Correlation, Item-test Correlation, KR-21); (3) Principal Components (Percent variance, Number of Eigenvalues > 1, Eigenvalue 1/Eigenvalue 2, Eigenvalue/variance, Sum of Residuals, No. Residuals > .01, Correlation of raw & factor scores); (4) Factor Analysis (Chi-square 1 factor, Chi-square 2 factor-1 factor, Tucker & Lewis, Theta, Omega, Communalities, Non-linear factor analysis), and (5) Latent Trait Models (Christoffersson-Muthen, Two-Parameter Latent Trait Model, One-Parameter Latent Trait Model, Lord's chi-square); where factor analysis is the most widely used method (Ziegler & Hagemann, 2015). Dimensional analysis to determine the constructs that make up the instrument as one of the requirements before the analysis is often ignored (Ziegler & Hagemann, 2015). The concept of unidimensionality is important in testing for several reasons. First, varying abilities should not influence the measure of the construct on which the assessment is based, as occurs in the case of multidimensional measures; second, when ranking testees based on a construct, we only need to have one dimension so as not to confound the ranking of testees based on the quality being measured; third, the requirement of a unidimensional assumption is necessary to avoid potential bias in estimating item parameters and person parameters (Stout, 1987; Reckase, 1979; Harrison, 1986). Therefore, violations of the unidimensional assumption can result in the inability of the model to describe the data (Eleje et al., 2018), thereby misleading the testee's ability to measure results (Bichi & Talib, 2018). Based on this explanation, this research was conducted to prove the unidimensionality of the cognitive instrument for evaluating science learning, which was specifically carried out using Exploratory Factor Analysis (EFA) with the Principal Component Analysis (PCA) Extraction Method.

METHODS

The Cognitive Test Instrument of Evaluation Science Learning (CTIESL) is in the form of a description question totaling 13 items and was developed based on the stages of developing a test instrument by Mardapi (2008), which consists of nine steps, namely: (1) compile test specifications (2) write test questions (3) review test questions (4) conduct test trials (5) analyze test items (6) improve tests (7) assemble tests (8) carry out tests (9) interpret test results. However, in this article, the focus of the development steps is only up to step five because proving the dimensionality of the cognitive test construct of science learning evaluation is in that step. The test instrument development procedure must be followed to produce a good test instrument while answering the challenges that may be faced by test developers, especially those related to the length of the test, the number of criteria used to scale the responses given by testees, and the sample size on item standardization which also determines the level of stability achieved (Mardapi & Kartowagiran, 2011).

Each step is described as follows: The first step was to develop a test specification (see Table 1) with Bloom's Taxonomy distribution: C1 (item number 1, 2, 3, 4), C3 (item number 13), C4 (item number 5), C5 (item number 6), and C6 (item number 7, 8, 9, 10, 11, 12). The second step was to compile the test items through essay questions and scoring rubrics. Essays were chosen because this form of questioning supports students' use

of interpretation and understanding of the course (Dahl et al., 2023), and to respond to this form of questioning requires a combination of research skills, writing ability, critical thinking, and ethical knowledge (Malik et al., 2023), thus making it a valuable tool for learning and assessment (Noorozi et al., 2023; Haro et al., 2023). Scouller (1998) adds that students will learn more deeply when responding to essay questions that they perceive as assessing higher levels of cognitive processing. However, while there are weaknesses in essay questions, such as the possibility of subjective judgment, they can be overcome with scoring rubrics that help ensure a consistent approach when multiple raters are responsible for the same assessment (Chan & Ho, 2019; Sulla, 2011).

Table 1. CTIESL Specification Table

Subject	Item Number
Tests, Measurements, Assessments, and Evaluations	1, 2, 3, 4, 5
Validity and Reliability	6
Preparation of cognitive test items	7, 8, 9, 10, 11, 12
Preparation of test and non-test instruments	13

To overcome the weakness of essay questions and ensure the reliability of CTIESL, an analytical rubric was also developed as a scoring guide. The final score of the analytic rubric is the total of each rubric section. Each criterion, or rubric line, provides detailed feedback on students' strengths and weaknesses and helps students understand how well they demonstrate their learning of a particular skill or concept and what needs improvement (Suskie, 2018). In general, for students, rubrics help clarify the depth of understanding of specific components of an assignment and allow students to understand what they have learned and what they still need to know, while for teachers, rubrics will inform how well students are achieving the components of the assignment that are aligned with the learning outcomes. This information enables analysis of the appropriateness of tasks, resources, instruction, and other course-related elements that support learning of the outcomes taught in the course (Olson & Krysiak, 2021). The third step is to review the test questions through content validation by five experts who are processed through the following Aiken formula (1985).

Based on the validation data obtained and further processed through Aiken's formula, it can be concluded that the CTIESL items developed are content valid. This is evidenced by the distribution of the content validity index as follows: 0.86 (item numbers 2, 7, 13), 0.90 (item numbers 1, 3, 4, 5, 11), 0.95 (item numbers 8, 9), and 1.00 (item numbers 6, 10, 12). The fourth step was to test the CTIESL instrument, which was responded to by all prospective science teacher students who took the Science Learning Evaluation course (98 people) in the 2023/2024 academic year. Finally, the fifth step is to analyze the test items through quantification of the CTIESL response results using the assessment rubric guide so that the polynomial data is obtained and the unidimensionality of CTIESL can be proven through Exploratory Factor Analysis (EFA) with the Principal Component Analysis (PCA) Extraction Method using IBM SPSS 25 software.

RESULT AND DISCUSSION

Concerning the analysis of test instruments, determining the number of dimensions underlying an item score is the most fundamental problem in assessing psychometric properties (Ziegler & Hagemann, 2015) because the correct specification of the number of latent dimensions is directly related to the construct validity of the test (van Rijn et al., 2016) as well as being the basis for determining appropriate IRT analysis software and methods (Mitee, 2019); and Factor Analysis (FA) is by far the most widely used tool (Izquierdo et al., 2015; Muñiz & Fonseca-Pedrero, 2019). Exploratory Factor Analysis is included in a multivariate statistical method of Factor Analysis (Williams et al., 2010) which attempts to identify the smallest number of hypothetical constructs (also known as factors, dimensions, latent variables, synthetic variables, or internal attributes) that can explain observed covariation between a set of measured variables (also called observed variables, manifest variables, effect indicators, reflective indicators, or surface attributes) (Watkins, 2018); so it is widely applied and widely promoted to develop and refine instruments, clinical assessments to produce evidence of construct validity (Tavakol & Wetzel, 2020) especially for researchers who do not have a stable hypothesis about the structure of the instrument (Kline, 2023).

A CTIESL has been developed in this research, and proof of its unidimensionality is carried out through EFA using the Principal Component Analysis (PCA) extraction method. The PCA extraction method was chosen because this method is most widely used as the default method in many statistical software (Thompson, 2004), does not require a theoretical basis or previous model (Gorsuch, 1983), and is recommended in establishing an initial solution for EFA (Pett et al., 2003). According to Costello & Osborne (2005), Factor Analysis is preferable to Principal Component Analysis (PCA), which is only a data reduction approach. If the researcher initially develops an instrument with several items and is interested in reducing the number of items, then PCA will provide maximum results (Netemeyer et al., 2003). It should be understood that PCA as an extraction method

in EFA and PCA as a method used to evaluate the number of dimensions based on [Hattie \(1985\)](#) are two different things. The same "PCA" terminology can indeed mislead previous analyses because it seems as if they share standard features ([Santoso et al., 2022](#)). The initial step in proving the unidimensionality of the CTIESL begins with ensuring sample adequacy and ensuring that the variable matrix being tested is not orthogonal (correlated enough so that the correlation matrix deviates significantly from the identity matrix). Sample adequacy was tested through the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO-MSA) and anti-image correlation matrix, where both (Table 2 and Table 3) have met the KMO-MSA cut-off value (higher than 0.5), and the cut-off value anti-image correlation matrix (higher than 0.5). Therefore, it can be concluded that the data set can be reduced to its factor structure with EFA. Meanwhile, Bartlett's Test of Sphericity attempts to reject ($\text{sig} < 0.05$) the null hypothesis (H_0), which states that the orthogonal variables or original correlation matrix is an identity matrix so that it can show that the variables are not related and, therefore, are not suitable for structure detection ([Shrestha, 2021](#)). Based on these criteria, Bartlett's Test of Sphericity is also significant and fulfilled (Table 2), so the factor analysis can be continued.

Table 2. KMO-MSA and Bartlett's test of sphericity

Statistics		Test Results
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		,878
Bartlett's Test of Sphericity	Approx. Chi-Square	1425.981
	df	78
	Sig.	,000

Table 3. Anti-image correlation matrix

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃
X ₁	,842 ^a	-.407	-.243	-.047	.107	-.054	-.183	-.228	.111	,059	-.713	,458	,195
X ₂	-.407	,891 ^a	-.139	-.150	,129	-.092	,007	-.072	,007	-.202	,411	-.594	-.017
X ₃	-.243	-.139	,878 ^a	,255	,159	-.382	-.207	.125	-.167	-.088	,034	,236	-.540
X ₄	-.047	-.150	,255	,921 ^a	-.010	-.547	-.185	-.172	-.092	-.206	-.194	,215	.018
X ₅	.107	,129	,159	-.010	,879 ^a	-.585	-.304	-.371	,215	-.133	,053	,100	-.293
X ₆	-.054	-.092	-.382	-.547	-.585	,868 ^a	,115	,360	-.174	,100	-.022	-.088	,206
X ₇	-.183	,007	-.207	-.185	-.304	,115	,947 ^a	-.088	-.205	,058	,187	-.255	,027
X ₈	-.228	-.072	.125	-.172	-.371	,360	-.088	,905 ^a	-.226	-.264	-.066	,059	,095
X ₉	.111	,007	-.167	-.092	,215	-.174	-.205	-.226	,936 ^a	-.064	.128	-.290	,120
X ₁₀	,059	-.202	-.088	-.206	-.133	,100	,058	-.264	-.064	,936 ^a	,258	-.173	-.105
X ₁₁	-.713	,411	,034	-.194	,053	-.022	,187	-.066	.128	,258	,771 ^a	-.706	-.375
X ₁₂	,458	-.594	,236	,215	,100	-.088	-.255	,059	-.290	-.173	-.706	,796 ^a	-.032
X ₁₃	,195	-.017	-.540	.018	-.293	,206	,027	,095	,120	-.105	-.375	-.032	,863 ^a

Note: ^a = Measure of Sampling Adequacy (MSA)

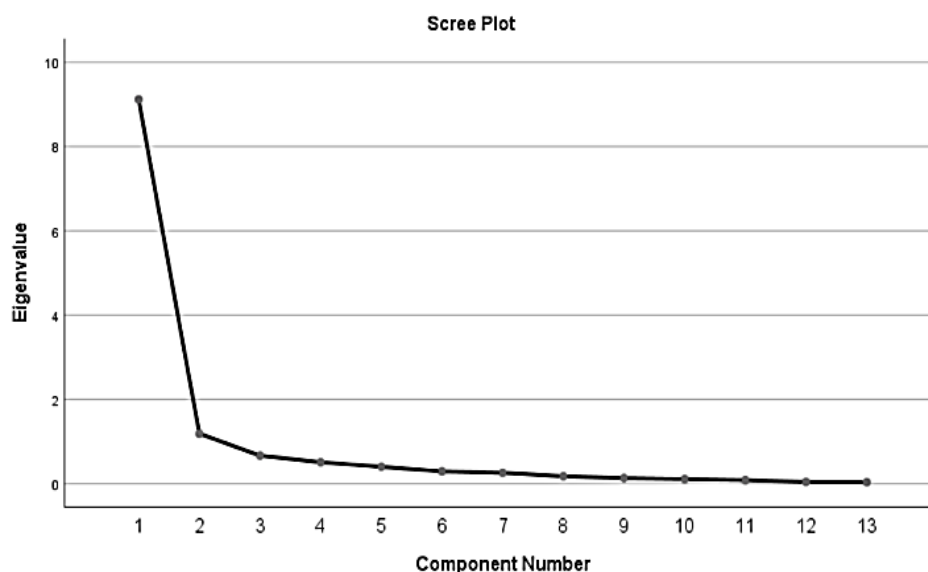
The next step is to extract data using the PCA extraction method, which aims to reduce many items and convert them into a collection of factors to obtain a unidimensional scale ([Williams et al., 2010](#)). Several criteria are better used simultaneously as considerations for making the right decision, for example, Kaiser's criterion (eigenvalue > 1) ([Kaiser, 1960](#)), the Scree test or Scree Plot ([Cattell, 1966](#)), and the cumulative percent of variance extracted ([Hair et al., 2010](#)). If we refer to these three criteria, the CTIESL instrument prepared is proven to be bidimensional based on Kaiser's criteria (first eigenvalue = 9.113 and second eigenvalue = 1.187) but is proven to be unidimensional based on the Scree Test criteria and the cumulative percent of variance extracted. The explanation for this conclusion is as follows.

First, bidimensionality is reviewed using Kaiser's criteria (Table 4). Kaiser's method states that only constructs with eigenvalues greater than one should be retained for interpretation ([Kaiser, 1960](#)). Initially, this method was the best known and had the easiest theoretical basis and application ([Fabrigar et al., 1999](#); [Gorsuch, 1983](#)). However, experts are starting to find several weaknesses; for example, this rule is quite arbitrary due to the lack of clarity in distinguishing factors with eigenvalues above one and below one ([Fabrigar et al., 1999](#); [Hayton et al., 2004](#)), apart from that this approach also shows a tendency to overestimate the number of factors ([Linn, 1968](#)) and in some cases, even underestimate it ([Horn, 1965](#); [Zwick & Velicer, 1986](#)). Ultimately, some studies suggest that the Kaiser Method is one of the least accurate methods for selecting retention factors ([Velicer & Jackson, 1990](#); [Fabrigar et al., 1999](#); [Ledesma & Valero-Mora, 2007](#)).

Table 4. Total variance explained using the PCA extraction method

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% Variance	Cumulative %	Total	% Variance	Cumulative %
1	9,113	70,100	70,100	5,717	43,976	43,976
2	1,187	9,131	79,232	4,583	35,255	79,232
3	,665	5,112	84,343			
4	,508	3,906	88,249			
5	,402	3,090	91,340			
6	,291	2,238	93,578			
7	,259	1,990	95,567			
8	,176	1,350	96,918			
9	,133	1,019	97,937			
10	,109	,835	98,772			
11	,083	,642	99,414			
12	,042	,322	99,736			
13	,034	,264	100,000			

Second, unidimensionality in terms of the Scree Plot (Figure 1). The Scree Plot method involves visual exploration of graphical representations of eigenvalues for breaks or discontinuities (Cattell, 1966) so that with this method, the number of points (plots) before the elbows on the scree plot (excluding points that become elbows) is the number of factors worth retaining. The logic behind this method is that the points will divide the factors into two: major and minor (Ledesma & Valero-Mora, 2007). Although the interpretation of scree plots can be subjective (Gorsuch, 1983; Tabachnick & Fidell, 2001; Thompson, 2004) so, the number of factors that need to be considered and the results may differ (Zwick & Velicer, 1986; Pett et al., 2003), the Scree Plot has good performance and more precise than the Kaiser method (Zwick & Velicer, 1986) and is a good choice for researchers (Costello & Osborne, 2005).

**Figure 1.** The unidimensionality of the CTIESL based on the scree plot

Finally, unidimensionality is viewed from the cumulative percentage of variance (Table 4). The cumulative percentage of variance method is based on achieving a certain cumulative percentage of the total variance extracted by successive factors so that the practical significance of the derived factors in explaining a certain amount of variance can be ensured (Hair et al., 2010). No absolute threshold is adopted for all applications (Henson & Roberts, 2006). However, in natural science research, the factoring procedure should usually not be stopped until the extracted factors account for at least 95% of the variance or until the last factor accounts for only a small portion (less than 5%), and in the social sciences, it accounts for 60% of the total variance (sometimes even smaller) is considered satisfactory (Hair et al., 2010). Ultimately, although there are strong reasons for unidimensionality, it must be emphasized that this property is not absolute but relative. That is, relative to the variable definition, perspective, measurement objectives, score interpretation, and frame of reference used. Considering broad (multidimensional) definitions of variables and constructs may be useful and desirable in certain situations. In contrast, more focused and narrow constructs would be more appropriate in

other situations (Hagell, 2014) because the consequences would be fatal if unidimensional scores were treated multidimensionally and vice versa (Garrido et al., 2019). Another point that must also be considered is that even in multidimensional measurement models, the scores produced in each dimension and the overall score summary must still be unidimensional within their respective reference frames and according to their respective definitions (Hagell, 2014).

To refine the measure, EFA allows researchers to explore factor results and other criteria (e.g., prior theory, minimum partial mean, parallel analysis, conceptual meaningfulness, etc.) (Tavakol & Wetzel, 2020). Research into the development of measurement instruments that do not apply the extraction, rotation, and interpretation recommended in EFA does run the risk of producing incorrect conclusions regarding the validity of a measure. Therefore, by developing an instrument through appropriate development steps, following construct analysis procedures (i.e., Exploratory Factor Analysis and/or Confirmatory Factor Analysis, CFA), a snapshot of the statistical relationships of key behaviors, attitudes, and dispositions of the construct of interest will be obtained. This snapshot provides important evidence for measurement validity based on the congruence of the test content with the construct's underlying theoretical framework. Furthermore, the inter-factor relationships that can be explored with EFA and confirmed with CFA can help researchers interpret the theoretical relationships between the underlying dimensions of a construct, thereby implicating measurement accuracy following the stated measurement objectives.

CONCLUSION

Determining the number of dimensions underlying an item score is one of the most fundamental problems in assessing the psychometric characteristics of an instrument, where Factor Analysis is the most widely used tool to overcome this problem. EFA with the PCA Extraction Method concludes that the CTIESL has a unidimensional construct based on the Scree Test criteria (detected only one point before the elbow on the Scree plot). The cumulative percent of variance extracted shows a quite satisfactory value (79.232%), so the instrument is proven only to measure one latent trait, namely the cognitive ability to evaluate science learning.

REFERENCES

- Aiken, L.R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and psychological measurement*, 45(1), 131-142.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole
- Ayanwale, M. A., Chere-Masopha, J., & Morena, M.C. (2022). The classical test or item response measurement theory: The status of the framework at the Examination Council of Lesotho. *International Journal of Learning, Teaching and Educational Research*, 21 (8), 384-406
- Bichi, A. A., & Talib, R. (2018). Item response theory: An introduction to latent trait models for testing and item development. *International Journal of Evaluation and Research in Education*, 7 (2), 142-151
- Cappelleri, J.C., Lundy, J.J., & Hays, R.D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, 36 (5), 648-662
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1 (2), 245-276
- Chan, Z., & Ho, S. (2019). Good and Bad Practices in Rubrics: The Perspectives of Students and Educators. *Assessment & Evaluation in Higher Education*, 44(4), 533-545.
- Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10 (7), 1-9
- Dahl, B. M., Vasset, F., & Frilund, M. (2023). Students' approaches to scientific essay writing as an educational method in higher education: A mixed methods study. *Social Sciences & Humanities Open*, 7(1), 100389.
- Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational and Social Sciences*, 3 (1), 57-75
- Erguven, M. (2013). Two approaches to psychometric process: Classical test theory and item response theory. *Journal of Education*, 2 (2), 23-30.

- Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4 (3), 272-299.
- Garrido, C.C., González, D.N., Seva, U.L., & Piera, P.J.F. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor analytic approach for assessing the dimensionality of tests and items. *Psychothema*, 31 (4), 450-457.
- Glas, C. A. (2008). Item response theory in educational assessment and evaluation. *Mesure et évaluation en éducation*, 31 (2), 19-34.
- Gorsuch, R. (1983). *Factor Analysis*. Hillsdale Erlbaum.
- Hagell, P. (2014). Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. *Open Journal of Statistics*, 4 (6), 456-465.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Pearson Education Inc.
- Haro, A. V., Noroozi, O., Biemans, H. J., Mulder, M., & Banihashem, S. K. (2024). How does the type of online peer feedback influence feedback quality, argumentative essay writing quality, and domain-specific learning? *Interactive Learning Environments*, 32(9), 5459-5478.
- Harrison, D. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, 11, 91-115
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9 (2), 139-164.
- Hayton, J.C., Allen, D.G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7 (2), 191-205.
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comments on improved practice. *Educational and Psychological Measurement*, 66 (3), 393-416.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2) , 179-185.
- Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The integration of classical testing theory and item response theory. *Psychology*, 12, 1397-1409.
- Izquierdo, I., Olea, J., & Abad, F.J. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psychothema*, 26 , 395-400.
- Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20 (1), 141-151
- Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford Publications
- Ledesma, R.D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12 (2), 1-11
- Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika*, 33, 37-71
- Livingston, S. A. (2020). *Basic concepts of item response theory: A nonmathematical introduction research memorandum (ETS RM-20-06)*. Educational Testing Service
- Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., & Darwis, A. (2023). Exploring artificial intelligence in academic essay: higher education student's perspective. *International Journal of Educational Research Open*, 5, 100296.
- Mardapi, D. (2012). *Measurement of educational assessment and evaluation*. Nuha Medika
- Mardapi, D., & Kartowagiran, B. (2011). Pengembangan instrumen pengukur hasil belajar nirbias dan terskala baku. *Jurnal Penelitian dan Evaluasi Pendidikan*, 15(2), 326-341.
- McBride, N. L. (2001). *An item response theory analysis of the scales from the international personality item pool and the neo personality inventory-revised* [Doctoral dissertation, Virginia Tech].

- Mitee, T. L. (2019). Comparative study of classical test theory and item response theory using item analysis results of quantitative chemistry achievement test. *The African Journal of Behavioral and Scale Development Research*, 1 (1), 26-36
- Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test [Ten steps for test development]. *Psychothema*, 31 (1), 7-16.
- Netemeyer, R.G., Bearden, W.O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Sage Publications
- Ohiri, S. C., & Okoye, R. O. (2023). Application of classical test theory as linear modeling to test item development and analysis. *International Research Journal of Modernization in Engineering Technology and Science*, 5(1), 2152-2159.
- Olson, J. M., & Krysiak, R. (2021). Rubrics as tools for effective assessment of student learning and program quality. In *Curriculum Development and Online Instruction for the 21st Century* (pp. 173-200). IGI Global.
- Pasquali, L. (2003). *Psicometria: Theory dos testes na psicologia and na educação*. Petropolis Vozes.
- Pasquali, L. (2007). Validade dos testes psicológicos: Será possível reencontrar o caminho? *Psicologia: Theoria e Pesquisa*, 23, 99-107.
- Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage Publications Inc.
- Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230
- Rusch, T., Lowry, P.B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management*, 54 (2), 189-203
- Santoso, P. H., Istiyono, E., & Haryanto, H. (2022). Principal component analysis and exploratory factor analysis of the mechanical waves conceptual survey. *Indonesian Journal of Psychological and Educational Measurement*, 11 (2), 209-225
- Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher education*, 35(4), 453-472.
- Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9 (1), 4-11.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 52, 589-617
- Sulla, N. (2011). *Students taking charge: inside the learner—active, technology—infused classroom*. London: Routledge Taylor & Francis Group
- Suskie, L. (2018). *Assessing student learning: A common sense guide* (3rd ed.). Hoboken, NJ: Jossey-Bass.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon
- Tavakol, M., & Wetzel, A. (2020). Factor analysis: A means for theory and instrument development in support of construct validity. *International Journal of Medical Education*, 11, 245 -247.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington DC: American Psychological Association
- Toni, P. M., Romanelli, E. J., & Salvo, C. G. (2004). Análise da estrutura fatorial da Bateria Piaget Head. *Boletim de Psicologia*, 54 (120), 73-86
- van Rijn, P.W., Sinharay, S., Haberman, S.J., & Johnson, M.S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assessments in Education*, 4, 1-23.
- Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis - some further observations. *Multivariate Behavioral Research*, 25 (1), 97-114
- Vincent, W., & Shanmugam, S. K. S (2020). The role of classical test theory to determine the quality of classroom teaching test items. *Pedagogia: Journal of Education*, 9 (1), 5-34

- Wagbara, E.S., & Nna-Kue, N.L. (2024). Relevance of classical test theory in the assessment of learning in tertiary institutions in Nigeria. *British Journal of Education, Learning, and Development Psychology*, 7(2), 202-2013
- Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, 44 (3), 219-246
- Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, 8, 1-13
- Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*, 31(4), 231–237.
- Zwick, W.R., & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3), 432-442.