Pegem Journal of Education and Instruction, Vol. 15, No. 3 , 2025 (pp. 60-77)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Teachers' Tendency to Use Alternative Assessment Methods

Güneş Korkmaz[1], Hasan Fehmi Özdemir[2], Çiğdem Çalışkan Toraman[3]*, Çetin Toraman[4]

[1]İstanbul Medeniyet University, Department of Medical Education, İstanbul, Turkey

[2]Şırnak University, Graduate School of Education, Department of Educational Sciences, Şırnak, Türkiye

[3]Ministry of National Education, Edremit Science and Art Center, Balıkesir, Turkey

[4]Çanakkale Onsekiz Mart University, Department of Medical Education, Çanakkale, Turkey

Abstract

Our study developed a tool to determine teachers' tendency to use alternative assessment methods and examined variations based on sociodemographic factors. We surveyed 1022 teachers from primary to high schools, dividing the data into two sets of 511. The first set underwent exploratory factor analysis, item response theory analysis, and reliability assessments to establish the scale's validity. Confirmatory factor analysis on the second set validated a three-factor construct. Analyzing the comprehensive dataset revealed trends in alternative assessment use concerning gender and seniority. Findings showed frequent use of concept mapping, observation, and interviews, with overall positive tendencies toward alternative assessments. Male teachers perceived these methods as less practical, while female teachers saw them as supportive of learner-centered education. Additionally, teachers with more seniority were less likely to use alternative assessment methods compared to those with 0-5 years of experience.

Keywords: Teachers, alternative assessment, classroom assessment, scale development, classical test theory, item response theory

Introduction

Assessment encompasses various aspects, including measuring students' achievement, defining student performance, summarizing what students can do, inferring what students could potentially do, and reflecting on the curriculum in action (Nasri et al., 2010). Therefore, as Herman (1992) states, assessment is not an end in itself but is a systematic process that facilitates teachers' instruction and students' learning. As constructivist theory and related educational philosophies embrace the diverse needs of today's learners (Yurdabakan, 2011), dissatisfaction with existing standardized testing has given rise to proposals for new assessment alternatives (Herman, 1992), the concept of *alternative assessment* has been a common topic of discussion. Gronlund (1998) states that alternative assessment is a method that provides an alternative to the traditional paper-and-pencil tests. Alternative assessment is considered to be a broad term that encompasses any methods of standardized multiple-choice testing that is commonly used in traditional educational environments (Gipps & Stobart, 2003), and researchers and teachers as practitioners have increasingly been trying to explore alternative assessment methods that go beyond the confines of traditional assessment methods. In this respect, assessment takes on a constructive approach, aiming to help students improve at their own pace rather than pass judgment. This highlights that each student's achievement is unique to them, rather than comparing them to other learners in the classroom (Buhagiar, 2007). Moreover, educators should use alternative assessment as students do not learn in the same way; thus, they cannot be assessed only in a uniform manner (Brualdi, 1996), and alternative assessment can support learners on various "real-life" skills such as critical thinking, reflective thinking, problem-solving and decision-making to prepare students for their future education steps and/or careers (Orakçı, 2021; Yurdabakan, 2011; Ayyoub et al., 2022; Ayaz & Gök, 2022). Some of the alternative assessment techniques are portfolio assessment, observation, self-assessment and peer assessment, concept mapping, diagnostic tree model, simulations, and written reflections (Ahmad, Sultana & Jamil, 2020; Corcoran,

Corresponding Author e-mail: cigdemcaliskantoraman@gmail.com

https://orcid.org/0000-0002-5862-628X

How to cite this article: Korkmaz G, Özdemir HF, Toraman ÇÇ, Toraman Ç. Teachers' Tendency to Use Alternative Assessment Methods. Pegem Journal of Education and Instruction, Vol. 15, No. 3, 2025, 60-77

Source of support: Nil

Conflicts of Interest: None.

DOI: 10.47750/pegegog.15.03.08

Received: 19.12.2024

Accepted: 02.03.2025 **Published:** 01.07.2025

Dershimer & Tichenor, 2004; Liu, Kitto & Shum, 2021; Nasri et al., 2010; Mofolo, 2023).

The shift towards alternative assessment has led to a transition from *assessment of learning* to *assessment for learning*, emphasizing that assessment serves not only summative purposes but also formative functions. *Assessment for learning* occurs when teachers use insights about students' progress to inform their teaching. This is contrasted with the *assessment of learning*, which evaluates what students have learned up to a certain point (Stacey et al., 2009). From this perspective, summative assessments, conducted at the end of instruction, are distinct from the educational process and aim to determine if students have acquired sufficient knowledge and skills over a specific period. In contrast, *assessment for learning* is integrated into the educational process, providing rich information to enhance each learner's progress (Schuwirth & Van der Vleuten, 2011). In this context, summative assessment and formative assessment are two related concepts. Summative assessment focuses on measuring and reporting students' learning and achievement at the end of a learning period, often using traditional methods like multiple-choice tests or standardized tests (Herman, 1992). Formative assessment, on the other hand, emphasizes continuous monitoring and gathering evidence of learners' progress throughout the learning and teaching process (Yan et al., 2021).

Considering the above-mentioned information, we can suggest that moving from *traditional assessment* to *alternative assessment* is not merely a pedagogical trend, but also a dynamic response to the call for a more holistic and nuanced understanding of learners' capabilities, and an indicator of our curriculum goals and instructional quality. It is discussed in the literature that, to ensure the development of students' learning, alternative assessment methods should be used and students should be allowed to improve their own learning (Cowan & Cherry, 2012; Gibbs, 2019). Furthermore, it should be kept in mind that the success of alternative assessment depends on how teachers perceive and implement formative assessment techniques during instruction (Yan et al., 2021). In other words, how teachers perceive their role in the assessment process and their willingness to use alternative assessment in their classrooms play an important role in classroom practice and successful implementation (Culbertson & Wenfan, 2003). Therefore, in this study, we aimed to develop a measurement tool that will determine teachers' tendency to use alternative assessment methods and to make comparisons about teachers' use of alternative assessment tools based on some socio-demographic variables. In line with this aim, the following are the research questions of our study:

1. What is the validity and reliability of the "Teachers' Tendency to Use Alternative Assessment Scale (T-TUAS)" based on Item Response Theory (IRT) and Classical Test Theory (CTT)?
2. What is teachers' tendency to use alternative assessment methods?
3. What are the most common alternative assessment techniques used by the teachers?
4. Does the level of teachers' tendency to use alternative assessment methods differ in terms of their gender and year of seniority?

Method

The process of developing a scale follows intricate and systematic steps, requiring accuracy in theory as well as methodology. According to Morgado et al. (2017), this process typically consists of three basic stages: (1) item generation which comprises developing an item pool; (2) theoretical analysis which assesses the new scale's content validity; and (3) psychometric analysis which focuses on evaluating the construct validity and reliability of the new scale. Evaluating construct validity can be accomplished through techniques like Exploratory Factor Analysis or Confirmatory Factor Analysis. In our study, we followed a similar approach. Various measurement tools can be devised utilizing different theoretical frameworks. Classical Test Theory (CTT) stands out as one of the most widely adopted scale development theories globally due to its straightforward and practical process. However, it's important to note that CTT comes with its limitations. For instance, the psychometric characteristics of a measurement tool based on CTT relies on the specific target group to whom the tool is applied. Additionally, just one standard error value can be obtained for an entire group when using CTT to design new measurement tools. On the contrary, Item Response Theory (IRT), also known as Modern Test Theory, operates differently. Item parameters are independent of the group of respondents in IRT, and the group's characteristics are independent of the items (Embretson & Reise, 2000). Hambleton et al. (1991) states that using IRT makes it possible to estimate a unique standard error for each participant, and to improve measurement precision.

Participants

Şırnak University Scientific Research Ethics Committee gave its consent for this study to be carried out (Date of Approval: 07/12/2023-No: 2023/82405). Additionally, permission for the application was received from Şırnak Governorship Provincial Directorate of National Education to provide the measurement tool to teachers (Date: 20/12/2023-No: E-61543340-604.01.01-92552958). The research was

Table 1: Distribution of the Teachers According To Sociodemographic Variables

Variables n(%)		Whole dataset n(%)	EFA Dataset n(%)	CFA Dataset
Gender	Female	580(56.8)	279(54.6)	301(58.9)
	Male	428(41.9	224(43.8)	204(39.9)
	Prefer not to answer	14(1.4)	8(1.6)	6(1.2)
Year of Seniority	0-5 Years	659(64.5)	344(67.3)	315(61.6)
	6-10 Years	230(22.5)	102(20)	128(25)
	11-15 Years	79(7.7)	34(6.7)	45(8.9)
	16-20 Years	32(3.1)	19(3.7)	13(2.5)
	21 Years and above	22(2.2)	12(2.3)	10(2)
Total		1022(100)	511(100)	511(100)

conducted with 1022 teachers. Using the random division approach, the entire dataset was divided into two halves. A random selection of 50% of the total data (511 teachers) was used for exploratory factor analysis (EFA), along with McDonald's omega, Cronbach's Alpha reliability coefficients, and IRT analysis. Confirmatory factor analysis (CFA) was also performed using the second half of the data (50% of the whole data, 511 teachers). According to Hair et al. (2014), CFA results can be verified in two different samples: one produced using a new application, or the other from a split sample of the original dataset.

Table 1 shows the descriptive statistics obtained from the teachers' data from the entire dataset, EFA and CFA.

Table 1 indicates that the majority of participants are women (580 [56.8%]), the majority of them have a seniority of 0-5 years (659 [64.5%]).

Preparing the draft of ''Teachers' Tendency to Use Alternative Assessment Scale (T-TUAS)" before the pilot application

The initial form of the "Teachers' Tendency to Use Alternative Assessment Scale (T-TUAS)" was prepared using the following procedure:

1. Determine the purpose of the scale: The purpose is to assess the level of teacher tendency to use alternative assessment methods.
2. Select the type of item: A 5-point Likert type was selected (The response categories are between Strongly disagree, disagree, partially agree, agree, and on the other end strongly agree).
3. Review the literature and compile a list of draft items: The literature on in-class assessment, measurement and evaluation, alternative assessment, performance evaluation, formative evaluation, summative evaluation, and teachers' use of in-class measurement and evaluation methods was reviewed and a candidate item pool was generated.
4. Obtain professional feedback on the proposed item pool: The initial item pool was sent to seven measurement and evaluation experts. The items that these experts deemed appropriate were added to the original form, while the items that they deemed inappropriate were taken out.
5. Make a linguistic redaction: To guarantee linguistic eligibility, a redaction in terms of language, expression and punctuation was taken from a Turkish language expert.

These procedures led to the creation of a draft scale form with 22 items and a 5-point Likert type. In this form, there were 10 reverse coded items about negative tendencies using alternative assessment methods.

Procedure

The following steps were followed in the scale development process:

1. The T-TUAS draft form was prepared.
2. The permission to conduct research about "Unveiling teachers' tendency to use alternative assessment methods" was obtained from Şırnak University Scientific Research Ethics Committee (Date of Approval: 07/12/2023-No: 2023/82405).
3. The research application permission was obtained from Şırnak Governorship Provincial Directorate of National Education (Date: 20/12/2023-No: E-61543340-604.01.01-92552958).
4. To create a hybrid application (online and in-person), both Google Forms and consent forms were prepared.
5. The authors visited some schools in Şırnak, Turkey and applied the scale face-to-face. Some of the data from the teachers were collected online.

6. Both online and in-person application data were integrated.
7. Item-total correlations based on Classical Test Theory, construct validity, and exploratory factor analysis were used to assess the validity of the scale. Furthermore, item and test information functions, item characteristic curves, item discrimination, and item difficulty levels were analyzed using IRT.
8. To assess the scale's reliability, for CTT Cronbach Alpha internal consistency, and McDonald's omega; and for IRT marginal reliability coefficients were calculated.

Data Analysis

Dataset Preparation

79 teachers' data were removed from the data set even though 1101 teachers took part in the survey. Because these 79 participants were unable to answer nearly half of the scale's items. The remaining 1022 teachers' data were imported into R and JAMOVI statistical programs. With the R "mvn" package, 22 items underwent multivariate normal distribution analysis utilizing the "Henze-Zirkler," "Mardia," and "Doornik-Hansen" approaches (Korkmaz et al., 2014). Multivariate normal distribution analysis was also applied to the exploratory factor analysis (EFA) data file consisting of 511 teachers' data and to the confirmatory factor analysis data file consisting of the other 511 teachers' data.

By comparing the correlation coefficient sizes with the partial correlation coefficient sizes in factor analysis, the Kaiser-Meyer-Olkin test (KMO) offers a gauge of sample adequacy (Pett, Lackey & Sullivan, 2003). The partial correlation coefficient is a measure of how strongly the items are related to each other. Following the removal of all other items' linear effects, partial correlations show the correlations between each pair of items (Hair et al., 2014; Pett, Lackey & Sullivan, 2003). A Kaiser-Meyer-Olkin test score of 0.90 or more is considered "excellent," 0.80 to 0.90 "good," 0.70 to 0.80 "acceptable," 0.60 to 0.70 "moderate," 0.50 to 0.60 "low level," and less than 0.50 "unacceptable," according to Kaiser and Rice (1974). Field (2018) states that a KMO score of at least 0.50 indicates the necessity for more data collection or a reevaluation of the variables that should be included in the study or analysis.

A correlation matrix in the form of an identity matrix is not desirable in factor analysis. The items in this matrix do not have a reciprocal relation; all correlation coefficients are near zero. Bartlett's Test of Sphericity already tests whether the data gathered is an identity matrix. The H_0 hypothesis—that is, the idea that there is no relation between the items—is examined in Bartlett's test of sphericity to determine whether the correlation matrix is an identity matrix. The H_0 hypothesis is rejected when a significant Bartlett test result is obtained (Field, 2018; Pett, Lackey & Sullivan, 2003). The Bartlett's Test of Sphericity is however significantly influenced by sample size, and in big samples, the p value is almost always significant (Field, 2018; Hair et al., 2014).

Exploratory Factor Analysis (EFA)

The "Principal Axis Factoring (PAF)" method was used in exploratory factor analysis. PAF analyzes by concentrating on shared variance rather than error sources specific to individual measurement. PAF models the shared variance in a set of n measurements and is more commonly employed in social and behavioral science fields (Warner, 2013).

To determine the number of factors during EFA, the eigenvalue was accepted as 1. When it comes to identifying latent roots, factors with eigenvalues higher than one are deemed significant. All factors with latent roots less than 1 are ignored as insignificant. Using the eigenvalue as a reference is reliable when there are 20 to 50 variables (number of items).

The outcome of the exploratory factor analysis is an explored construct. Divergent opinions exist regarding the proportion of the factor structure's variance that ought to be explained for the feature of interest, which in this case is the teachers' tendency to use alternative assessment methods. A result that explains 60% (even less in certain circumstances) of the explained total variance is considered sufficient in the social sciences, where knowledge is typically less precise, according to Hair et al. (2014). In a similar vein, Warner (2013) states that 40% to 70% should be the acceptable bounds. A significant factor load value is ±.50. A well-defined structure is indicated by ±.70 and above (Hair et al., 2014).

Confirmatory Factor Analysis (CFA)

In CFA analysis, the degree of model fit is important. By looking at the fit values, it is determined whether the model is suitable or not. The literature suggests that reference values for fit indexes determined for CFA are $0.05<RMSEA\leq0.08$ acceptable for RMSEA, $0>RMSEA\leq0.05$ excellent, 0.95 and above perfect for Tucker–Lewis index (TLI) and comparative fit index (CFI), $2< X^2/sd.$ for $X^2/sd. \leq5$ is considered acceptable, $0> X^2/sd\leq2$ is considered the perfect range (Anderson and Gerbing, 1984; Bentler, 1990; Hooper, Coughlan, and Mullen, 2008; Hu and Bentler, 1999; Kline, 2005; Marsh, Balla, and McDonald, 1988; Tabachnick and Fidell, 2013; Vieira, 2011).

Reliability Level of "Teachers' Tendency to Use Alternative Assessment Scale (T-TUAS)"

The degree of reliability of the constructed scale was evaluated using Cronbach's Alpha internal consistency coefficient,

McDonald's Omega reliability coefficient, and marginal reliability coefficient.

The alpha coefficient is a widely used method for calculating the internal consistency of polytomous items like items with Likert-type response categories. As the alpha value approaches 1, it is interpreted that there is high internal consistency among the items in the measurement tool, in other words, the measurement tool consists of homogeneous items that measure the same trait. In addition to its ease of calculation and interpretation, it also allows the calculation of the contribution of each item to reliability. Nunnally and Bernstein (1994) state that a reliability level of 0.70 or above is necessary.

The Omega (ω) coefficient developed by McDonald is one of the alternative reliability coefficients to Cronbach alpha. It is used when the covariances between the items are not equal, that is when the measurement tool consists of items with a congeneric structure (unequal factor loadings) and is calculated with the factor loadings of the items (McDonald, 1999). Similar to the alpha value, it takes a value between 0 and 1, and as this value approaches 1, it is interpreted as high internal consistency among the items.

The marginal reliability coefficient is a reliability coefficient estimated according to the IRT. According to Thissen (1991) and Flannery et al. (1995), it is the arithmetic mean of the reliability coefficients determined independently for various levels of the assessed psychological trait. This means that the reliability coefficient computed for the entire instrument is recognized as the marginal reliability coefficient. A high value of this coefficient is an indicator of the reliable results obtained from the applied instrument.

Item Response Theory (IRT) Analysis

Unidimensionality and local independence assumptions should be examined in reliability, and validity examinations with Item Response Theory (IRT) (Zhao, 2008). Unidimensionality requires that there is only one characteristic that affects the performance of individuals in the measurement tool, in other words, the relevant items of the measurement tool should be related to only one characteristic (Hambleton, Swaminathan & Rogers, 1991). EFA was preferred in this study to test the unidimensionality. The EFA results show that the "Teachers' Tendency to Use Alternative Assessment Scale (T-TUAS)" contains three subdimensions, or factors, as is mentioned under the following "Findings" section. In the study, IRT analyses were conducted taking into account each subdimension as a separate factor. Yen's (1993) Q3 statistic was used to test the local independence assumption, and the multidimensional item response theory (MIRT) (Chalmers, 2012) was used for the IRT calibrations.

According to IRT, an ideal discrimination value of an item should be between 0.5 and 2. It is also stated that this parameter is within the acceptable range when it is between 0.75 and 2.50 (Flannery, Reise & Widaman, 1995). The ideal limits for item difficulty levels are considered to be between -1.00 and 1.00. Items with a difficulty level below -1.00 and over 1.00 are accepted as easy on ability or achievement tests.

The item information function is a graphical display that indicates the range of the attribute (the trait that the scale seeks to measure) that the item most effectively uses to differentiate the individuals taking the measurement instrument (Edelen and Reeve, 2007). The test information function shows the extent to which the items of the instrument as a whole provide information about the trait of interest (Hambleton, Swaminathan, and Rogers, 1991).

Analyses Based on Sociodemographic Variables

Comparative analyses were conducted regarding teachers' tendency to use alternative assessment methods according to their gender, the faculty they graduated from, year of seniority, school type they work in, the types of settlements they work in, and whether they have attended any in-service training about alternative assessment methods.

Since the data set was very large, comparison analyses were performed with parametric techniques without examining the normal distribution and homogeneity of variance of the scores obtained from the Scale of Teachers' Tendency to Use Alternative Assessment Methods. The skewness value in large samples doesn't deviate substantially from normal distribution. In samples larger than 100, positive kurtosis starts to disappear and in samples larger than 200, negative kurtosis starts to disappear (Tabachnick & Fidell, 2013). In the interpretation of statistically significant relations, the limits proposed by Cohen (1988) for η^2 (eta-square); large effect size ≥0.14, medium effect size ≥0.06, and small effect size ≥0.01 were taken into consideration. For Cohen's d effect size, large effect size ≥0.8, medium effect size ≥0.5, and small effect size ≥0.2 limits were used.

Findings

CTT Validity Evidence of the ''Teachers' Tendency to Use Alternative Assessment Scale (T-TUAS"

The normal distribution of 22 items of the scale was examined in 3 data files: The complete data file of 1022 teachers, the exploratory factor analysis data file of 511 teachers, and the confirmatory factor analysis data file of 511 teachers. A multivariate normal distribution analysis was conducted

using R and the "mvn" package, employing the "Henze-Zirkler," "Mardia," and "Doornik-Hansen" methods. The results indicated that a multivariate normal distribution could not be achieved (p<.05).

The KMO value, which assesses the suitability of the data set for factor analysis based on the sizes of partial correlation coefficients, was also evaluated. For the EFA subset of 511 teachers, the KMO value was 0.893. Bartlett's Test of Sphericity, which checks if the correlation matrix resembles an identity matrix, yielded a value of 3,809 and was found to be significant (df= 91, p<.05). These results indicate that the data set was appropriate for factor analysis.

Item Total Correlation, Exploratory Factor Analysis (EFA), Factor Load Values of T-TUAS

Principal axis factoring (PAF) was utilized for the EFA factor analysis. In the EFA data set, the item-total correlation values and initial factor loadings for 22 items were examined. As a result of the analysis, 8 items (2, 12, 13, 14, 15, 16, 17, and 18) with correlation values less than .30 were removed from the draft pool. The initial, extraction, and item-total correlation values of the remaining 14 items were analyzed using Eigenvalue and Scree Plot Analysis to determine the number of factors in the scale. It was found that three factors had eigenvalues greater than 1. These three factors together explained 57.8% of the variance in the characteristic of interest (tendency to use alternative assessment methods). Therefore, it was concluded that the T-TUAS has a three-factor structure (Table 2).

Table 2.: T-TUAS Factors and Eigenvalues

Factor	*Eigenvalue*	*Variance*
1	3.23	23.1
2	2.61	18.6
3	2.52	18
Total		59.7

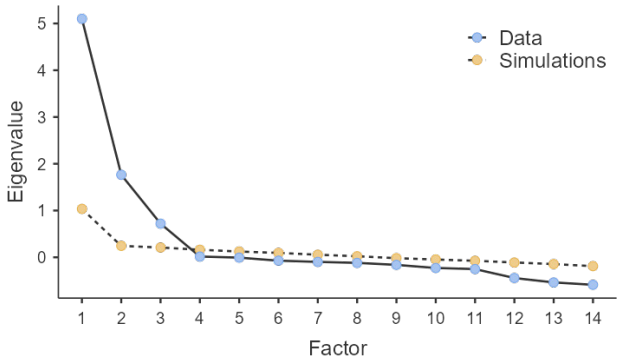


Figure 1: Scree plot from T-TUAS data in EFA

The eigenvalue of the first factor, which explains the highest variance (23%), is 3.23, the eigenvalue of the second factor is 2.61, and the eigenvalue of the third factor is 2.52. These factors explain 59.7% of teachers' tendency to use alternative assessment methods, which is acceptable according to the studies in the literature. The Scree Plot obtained from T-TUAS confirmed the three-factor structure (Figure 1).

Figure 1 shows that T-TUAS comprises three factors (each with an eigenvalue greater than 1). These results showed that T-TUAS has a three- factor structure. To identify which factors the remaining 14 items in the T-TUAS belonged to, axis rotation was performed. Given the interrelation among the three factors, Oblimin rotation was applied. Table 3 lists the items categorized under each factor.

As detailed in the data analysis section, the correlation values in Table 3 fall between the ranges recommended by the exploratory factor analysis (EFA) literature. It was found that 6 items were loaded onto factor 1, 4 items onto factor 2, and 4 items onto factor 3. Following the rotation process, factor load values ranged from .548 to .846. Table 4 presents the factors and their corresponding items.

Table 3: Factors After Oblimin Rotation and the Item Total Correlations

	Factor					
	1		*2*		*3*	
Items	*FLVR*	*ITC*	*FLVR*	*ITC*	*FLVR*	*ITC*
AA6 I recommend the use of alternative assessment methods to my colleagues.	0.823	0.677				
AA9 I use alternative assessment methods to obtain detailed information about students' learning.	0.750	0.676				
AA7 It is essential to use alternative assessment methods for students to internalize their acquisitions.	0.728	0.623				
AA5 I use alternative assessment methods because there is no method that can show student development better.	0.727	0.442				

	Factor					
	1		2		3	
Items	*FLVR*	*ITC*	*FLVR*	*ITC*	*FLVR*	*ITC*
AA1 I use alternative assessment methods because they are useful in determining learner outcomes.	0.633	0.570				
AA4 Alternative assessment method is a unique method of demonstrating learners' knowledge, skills and achievements.	0.584	0.570				
AA11 I do not use alternative assessment tools because they are difficult to prepare. (R)			0.820	0.432		
AA8 I do not use alternative assessment methods because it is a waste of time. (R)			0.809	0.484		
AA10 I do not use alternative assessment methods due to students' reluctance. (R)			0.786	0.442		
AA3 I do not use alternative assessment methods because they place a great burden on teachers. (R)			0.759	0.442		
AA21 Alternative assessment methods enable learners to take responsibility for their own learning.					0.846	0.641
AA20 Alternative assessment methods enable learners to participate in the assessment process.					0.838	0.620
AA22 Alternative assessment methods are effective in giving feedback to learners.					0.776	0.630
AA19 Alternative assessment methods provide effective feedback in developing learners' higher-order thinking skills.					0.548	0.406

(R) Reverse-Coded Items

FLVR: Factor Load Value after Rotation

ITC: Item Total Correlation

Table 4: T-TUAS factors and Item Information

Factors		*Items*	*Reverse coded items*	*The highest possible score*	*The lowest possible score*
Factor 1	Approach that finds alternative assessment methods useful (UAAM)	1, 4, 5, 6, 7 and 9	—	6	30
Factor 2	Approach that finds alternative assessment impractical (IAAM)	3, 8, 10 and 11	All	4	20
Factor 3	Approach that finds alternative assessment support learner-centered education (SAAM)	19, 20, 21 and 22	—	4	20

Confirmation of T-TUAS's factor structure

The result of the analysis revealed that T-TUAS has a 3-factor structure consisting of 14 Likert-type items. The validity of this structure was examined through confirmatory factor analysis (CFA), using data from 511 teachers. Figure 2 displays the diagram obtained from the CFA.

Figure 2 shows that the item with the lowest correlation under factor 3 is Item 19 (.62). Fit index values were found to be CFI= .970, TLI= .963, RMSEA=.056, X^2/df=2.607. These values suggest that the three-factor structure was confirmed.

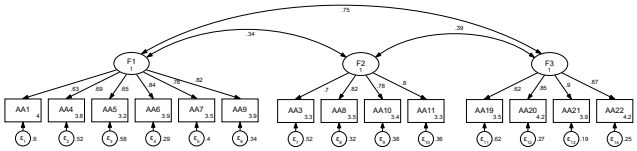


Fig. 2: CFA Diagram of T-TUAS's 3-Factor

T-TUAS's reliability evidence

T-TUAS reliability evidence was collected using Cronbach's Alpha internal consistency, McDonald's Omega

reliability, and Marjinal reliability coefficients. Both the 14 items were analyzed as a single factor scale and the 3 factors were subjected to separate reliability analyses. The results are presented in Table 5.

Table 5: T-TUAS's Cronbach's Alpha and McDonald's Omega Reliability Level

Scale or Sub-Factor	*Cronbach Alpha*	*McDonalds Omega*	*Marginal*
T-TUAS Scale Total Score	0.875	0.884	—
F1	0.872	0.874	0.884
F2	0.875	0.875	0.839
F3	0.851	0.860	0.865

The reliability values obtained for T-TUAS were 0.70 and above. According to Nunnally and Bernstein (1994), a satisfactory reliability level should be at least 0.70.

T-TUAS's Factor 1: Approach that finds alternative assessment 4methods useful (UAAM) IRT Validity Evidence

Item Response Theory analyses were conducted using the first half of the data set of 511 teachers. Item calibrations in IRT were performed using the Generalized Partial Credit Model (GPCM). The S_χ2 (degree of freedom), RMSEA, and level of significance statistics for the items according to GPCM are presented in Table 6.

Table 6: UAAM'S Item Fit-Indexes according to IRT

	GPCM				*GPCM*		
Item	*S_χ2*	*df*	*RMSEA*	*Item*	*S_χ2*	*df*	*RMSEA*
AA1	83.916	37	0.035	AA6	59.458	25	0.037
AA4	61.104	35	0.027	AA7	54.398	31	0.027
AA5	92.188	36	0.039	AA9	52.566	27	0.030

The boundary value for RMSEA, an essential fit index in Item Response Theory, is 0.08, with values below this threshold indicating item fit (Stout, 1990). According to the item fit statistics, the RMSEA values for all items were less than 0.039. Based on this result, it was determined that 6 items provided model fit according to the GPCM. In the next step, the "*a*" (item discrimination) and "*b*" (item difficulty) parameters, along with the standard errors for the items that fit the model according to GPCM, were estimated separately for each item. The results are presented in Table 7.

In IRT, an ideal scale item's discrimination value (parameter "*a*") should be between 0.5 and 2. However, the literature suggests that this value should be between 0.75 and 2.50. All items of UAAM, except for AA6, are ideal. The "a" parameter for nine items under the SAS factor ranged between 0.71 and 2.34. Estimates made according to the GPCM (LogLikelihood, p<.05) confirm the fitness of the measurement tool items. The item characteristic curves are shown in Figure 3.

Item characteristic curves show that items AA1, AA4, AA5, AA6, AA7 and AA9 are compatible with their options in the response set.

The graphical tool known as the item information function shows the range of attributes by which the item most successfully separates participants on the scale (Edelen & Reeve, 2007). In this function, a higher curve peak indicates greater informativeness of the items. Upon reviewing the item information functions of the UAAM items, it is found that items AA6, AA7, and AA9 are the most informative. Conversely, items AA1, AA4, and AA5 provide less information.

T-TUAS's Factor 2: Approach that finds alternative assessment impractical (IAAM) IRT Validity Evidence

As a result of item calibrations in IRT performed with the Generalized Partial Credit Model (GPCM), the level of significance statistics, S_χ2, (degree of freedom), RMSEA of the items are indicated in Table 8.

Table 7: UAAM's item parameters and standard error values according to GPCM

Items	**a(SE)**	**b1(SE)**	**b2(SE)**	**b3(SE)**	**b4(SE)**
AA1	1.235(0.089)	-2.197(0.217)	-2.050(0.144)	-0.415(0.076)	0.668(0.080)
AA4	1.354(0.095)	-2.134(0.180)	-1.776(0.118)	0.005(0.068)	1.247(0.091)
AA5	1.289(0.091)	-1.885(0.143)	-1.279(0.097)	0.503(0.077)	1.594(0.110)
AA6	3.045(0.247)	-2.002(0.111)	-1.402(0.071)	-0.289(0.046)	0.719(0.053)
AA7	1.920(0.136)	-2.031(0.134)	-1.409(0.084)	-0.095(0.056)	1.073(0.071)
AA9	2.314(0.168)	-2.265(0.147)	-1.516(0.082)	-0.261(0.051)	0.955(0.063)

Iteration=42 LogLikelihood: - 6939.821 p<.05

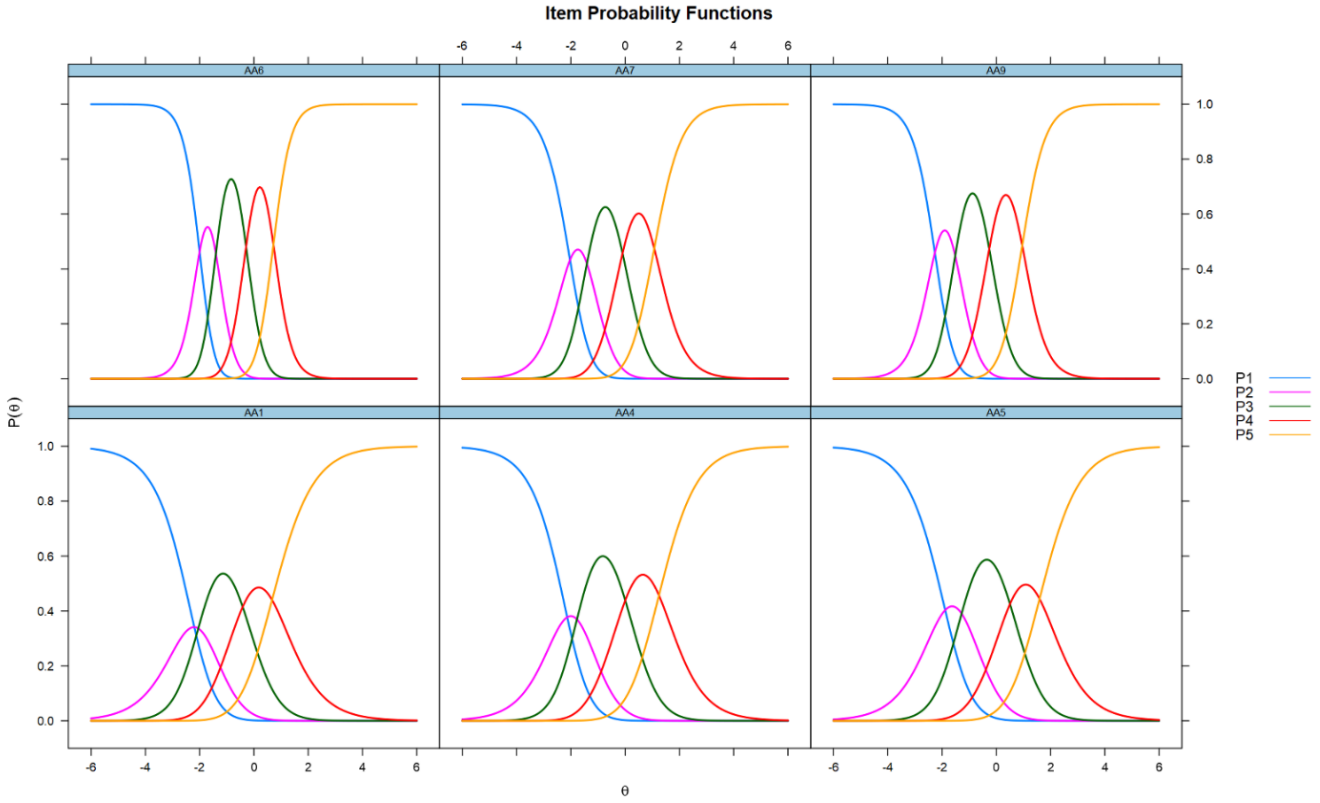


Fig. 3: Item characteristic curves of UAAM Items

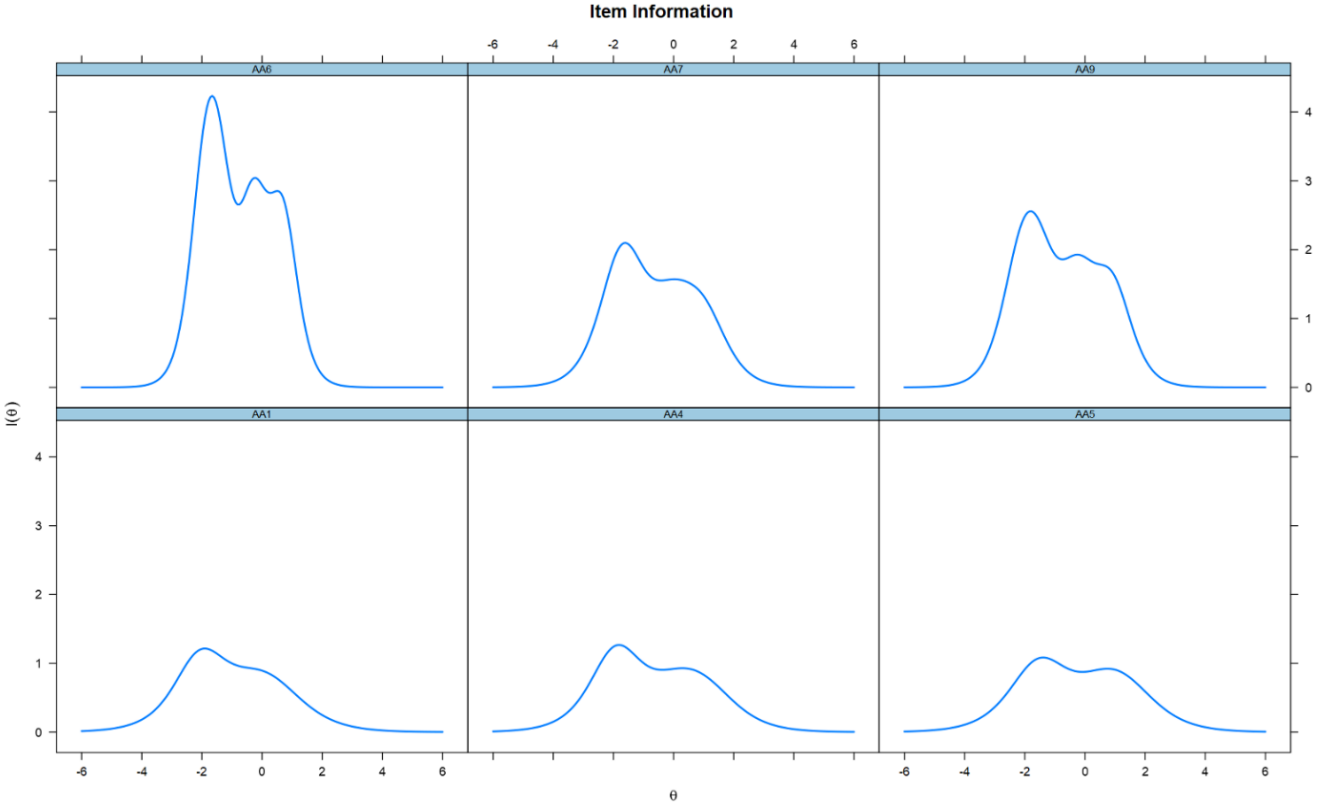


Fig. 4: Item information functions of UAAM

Table 8.: IAAM's Item Fit-Indexes according to IRT

	GPCM		
Item	*S_χ2*	*df*	*RMSEA*
AA3	94.686	21	0.059
AA8	64.913	18	0.051
AA10	85.571	19	0.059
AA11	77.883	20	0.053

Based on the item fit-index statistics, all items have RMSEA values below 0.059. Consequently, it was established that four items fit the model according to the Generalized Partial Credit Model (GPCM). Following this, the parameters for "*a*" (item discrimination) and "*b*" (item difficulty), along with their standard errors, were estimated for each item that showed model fit based on the Generalized Partial Credit Model. The results of these estimations are detailed in Table 9.

The analysis showed that all of the 4 items of IAAM are at the ideal discrimination levels. Estimations made according to the GPCM (LogLikelihood, p<.05) prove the concordance of the measurement tool items. Item characteristic curves are shown in Figure 5.

Figure 6 indicates item information functions.

Figure 6 shows that all items in IAAM items provide high level information.

T-TUAS Factor 3: Approach that finds alternative assessment support learner-centered education (SAAM) IRT Validity Evidence

As a result of item calibrations in IRT performed with the Generalized Partial Credit Model (GPCM), the level of significance statistics, S_χ2, (degree of freedom), RMSEA of the items is indicated in Table 10.

All items, except for item AA19, have RMSEA values less than 0.068, according to the item fit statistics. Item AA19 slightly exceeded the upper limit of fit indices. This result led to the conclusion that 4 SAAM items offered model fit based on the Generalized Partial Credit Model (GPCM). In the next step, the parameters for "*a*" (item discrimination) and "*b*" (item difficulty), along with their standard errors, were estimated separately for each item that demonstrated model fit based on the GPCM. The results of these estimations are presented in Table 11.

Table 9: IAAM's item parameters and standard error values according to GPCM

Item	*a(SE)*	*b1(SE)*	*b2(SE)*	*b3(SE)*	*b4(SE)*
AA3	1.552(0.119)	-0.295(0.069)	0.211(0.068)	1.467(0.105)	1.669(0.136)
AA8	2.288(0.209)	-0.052(0.054)	0.565(0.059)	1.195(0.079)	1.694(0.108)
AA10	2.100(0.168)	-0.605(0.061)	0.140(0.055)	1.219(0.078)	1.778(0.113)
AA11	2.049(0.167)	-0.480(0.060)	0.164(0.057)	1.221(0.081)	1.611(0.107)
Iteration=47		LogLikelihood: - 4895.911	p<.05		

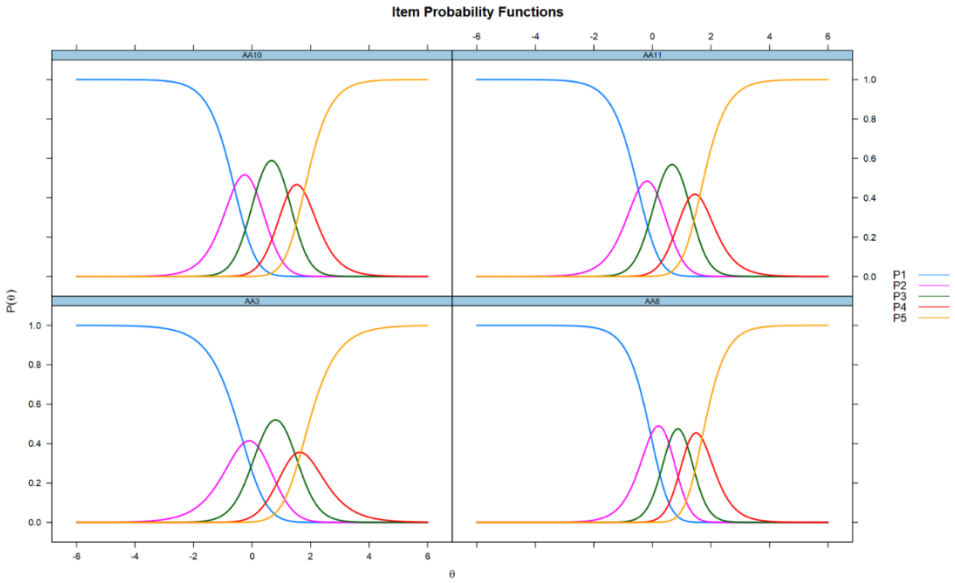


Fig. 5: Item characteristic curves of IAAM items

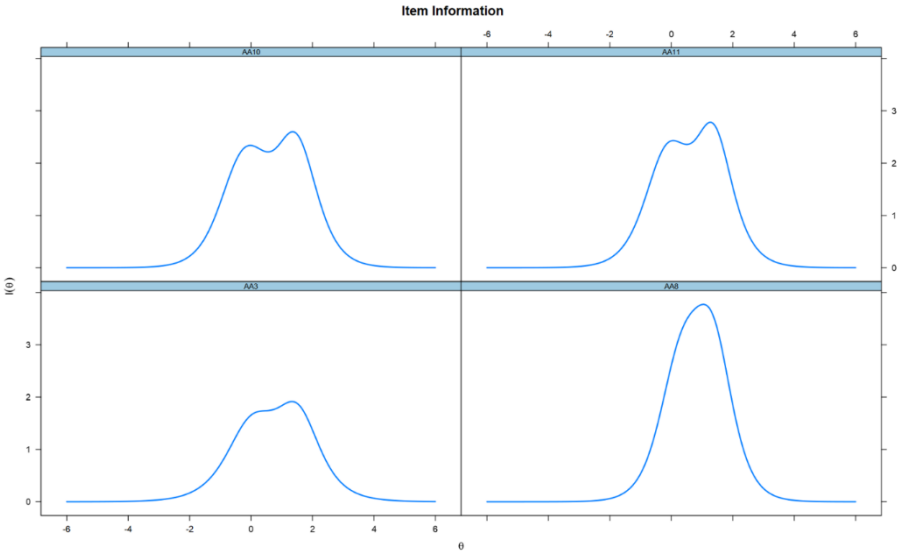


Fig. 6: Item information functions of IAAM

The analysis showed that all the items of SAAM are at the ideal level. Estimations according to the GPCM (LogLikelihood, p<.05) prove the concordance of the measurement tool items. Item characteristic curves are shown in Figure 7.

Item characteristic curves indicate that items AA20, AA21, and AA22 function well with their response options. The graph reveals that the "disagree" option for item AA19 does not align well with the other items in the response set. Item information functions are shown in Figure 8.

Table 10: *SAAM's Item Fit-Indexes according to IRT*

Item	GPCM		
	S_χ2	df	RMSEA
AA19	163.528	20	0.084
AA20	57.484	10	0.068
AA21	31.490	10	0.046
AA22	52.564	11	0.061

Table 11: SAAM's item parameters and standard error values according to GPCM

Item	*a(SE)*	*b1(SE)*	*b2(SE)*	*b3(SE)*	*b4(SE)*
AA19	0.970(0.071)	-2.093(0.214)	-1.827(0.147)	-0.172(0.090)	1.039(0.105)
AA20	3.390(0.269)	-2.136(0.129)	-1.560(0.075)	-0.510(0.046)	0.544(0.050)
AA21	4.258(0.379)	-2.026(0.108)	-1.435(0.066)	-0.464(0.043)	0.543(0.047)
AA22	3.202(0.242)	-2.210(0.141)	-1.623(0.080)	-0.536(0.047)	0.501(0.050)
Iteration=33	LogLikelihood: - 4336.919		p<.05		

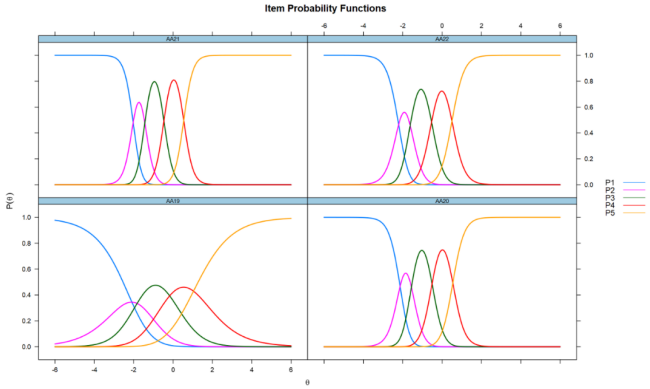


Figure 7. Item characteristic curves of SAAM items

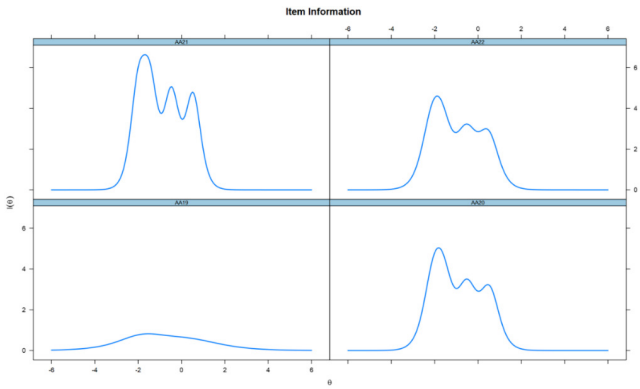


Fig. 8: Item information functions of SAAM

Item characteristic curves indicate items AA20, AA21, AA22 provide high level information, and the least informative is provided by item AA19.

Frequency of alternative assessment methods used by teachers

Apart from the T-TUAS scale items, teachers were asked 10 questions to obtain information about their frequency of use of alternative assessment methods and were asked to score them on a scale of 7 (never, almost never, rarely, sometimes, usually, often, always). The results are given in Table 12.

When the frequency of teachers' use of alternative assessment methods was examined, it was determined that the most used techniques were concept mapping, observation and interview. Other techniques are used at a moderate level.

Teachers' tendency levels to use alternative assessment methods

The responses of 1022 teachers to the T-TUAS items were examined with descriptive statistics. The results are given in Table 13.

Table 12: Frequency of teachers' use of alternative assessment techniques

Alternative Assessment Methods	*N*	*Mean (S. Dev.)*	*Median (Min. – Max.)*
How often do you use portfolio assessment?	1022	3.59(1.7)	4(1-7)
How often do you use project work in your assessment?	1022	3.85(1.6)	4(1-7)
How often do you use performance tasks as an alternative assessment?	1022	4.40(1.7)	4(1-7)
How often do you use checklists, rubrics (rating scales) to score performance tasks in your assessment??	1022	4.23(1.8)	4(1-7)
How often do you use self-assessment in your assessment?	1022	4.34(1.7)	4(1-7)
How often do you use peer-assessment in your assessment?	1022	3.87(1.8)	4(1-7)
How often do you use concept mapping technique in your assessment?	1022	4.51(1.7)	5(1-7)
How often do you use observation in your assessment??	1022	5.63(1.5)	6(1-7)
How often do you use interview in your assessment?	1022	5.11(1.6)	5(1-7)
How often do you use diagnostic tree model in your assessment?	1022	3.59(1.8)	4(1-7)

Table 13: Teachers' tendency levels to use alternative assessment methods

Items	*N*	*Mean (S. Dev.)*	*Median (Min.-Max.)*
AA1. I use alternative assessment methods because they are useful in determining learner outcomes.	1022	3.79(1)	4(1-5)
AA3. I do not use alternative assessment methods because they place a great burden on teachers. (*)	1022	2.26(1.2)	2(1-5)
AA4. Alternative assessment is a unique method of demonstrating learners' knowledge, skills and achievements.	1022	3.52(1)	4(1-5)
AA5. I use alternative assessment methods because there is no method that can show student development better.	1022	3.22(1)	3(1-5)
AA6. I recommend the use of alternative assessment methods to my colleagues.	1022	3.71(1)	4(1-5)
AA7. It is essential to use alternative assessment methods for students to internalize their acquisitions.	1022	3.54(1)	4(1-5)
AA8. I do not use alternative assessment methods because it is a waste of time. (*)	1022	2.06(1.2)	2(1-5)
AA9. I use alternative assessment methods to obtain detailed information about students' learning.	1022	3.66(1)	4(1-5)
AA10. I do not use alternative assessment methods due to students' reluctance. (*)	1022	2.38(1)	2(1-5)

Items	*N*	*Mean (S. Dev.)*	*Median (Min.-Max.)*
AA11. I do not use alternative assessment tools because they are difficult to prepare. (*)	1022	2.36(1.2)	2(1-5)
AA19. Alternative assessment methods provide effective feedback in developing learners' higher-order thinking skills.	1022	3.58(1)	4(1-5)
AA20. Alternative assessment methods enable learners to participate in the assessment process.	1022	3.87(1)	4(1-5)
AA21. Alternative assessment methods enable learners to take responsibility for their own learning.	1022	3.84(1)	4(1-5)
AA22. Alternative assessment methods are effective in giving feedback to learners.	1022	3.90(1)	4(1-5)

(*) Descriptive statistics were calculated without reverse coding. High scores indicate a negative tendency/approach to use alternative assessment methods.

Table 14: Teachers' tendency to use alternative assessment methods by gender (independent sample t-test)

Factors	*Gender*	*N*	*Mean (S. Dev.)*	*t*	*df*	*p*	*Cohen's d*
Approach that finds alternative assessment methods useful (UAAM)	Female	580	21.54(4.7)	0.676	1006	0.499	0.043
	Male	428	21.34(4.6)				
Approach that finds alternative assessment impractical (IAAM) (*)	Female	580	8.47(3.8)	5.750	1006	<0.0001	0.366
	Male	428	9.88(3.9)				
Approach that finds alternative assessment supporting for learner-centered education (SAAM)	Female	580	15.51(3.2)	3.197	1006	<0.0001	0.204
	Male	428	14.84(3.3)				
Tendency Level to Use Alternative Assessment Methods (**)	Female	580	52.57(9.2)	3.941	1006	<0.0001	0.251
	Male	428	50.31(8.7)				

(*) Descriptive statistics without reverse coding. High scores indicate a negative tendency/approach to use alternative assessment methods.

(**) After negative approach items were coded reversely, a total score was obtained. High scores indicate a positive tendency/approach to use alternative assessment methods.

When the teachers' responses to the T-TUAS scale items were examined, it was determined that the teachers agreed at a moderate level with the item "*I use alternative assessment methods because there is no method that can show student development better*". However, teachers' responses to the items "*I do not use alternative assessment methods because they place a great burden on teachers*", "*I do not use alternative assessment methods because it is a waste of time*", "*I do not use alternative assessment methods due to students*", "*I do not use alternative assessment tools because they are difficult to prepare*" are at a low level. It was determined that teachers mostly agreed with the other items on the scale. According to these results, it can be said that teachers, in general, have a positive tendency or approach to alternative assessment methods.

Teachers' tendency to use alternative assessment methods by gender

Teachers' tendency levels to use alternative assessment methods were compared according to their gender. This comparison was made using an independent sample t-test (the analysis was made only between male and female genders, 14 teachers who preferred not to answer about their gender were excluded as the number is not meaningful to make comparisons). The results are shown in Table 14.

According to the analysis results, there is no significant difference between female teachers and male teachers in terms of finding alternative assessment methods useful (UAAM) (p>.05). However, a significant relationship was noted between female teachers and male teachers' tendency that finds alternative assessment methods impractical (IAAM) (p <.05). This significant difference is at the small effect size level. Male teachers' tendency that finds the use of alternative assessment methods impractical (IAAM) is higher than female teachers' tendency.

A significant relationship was found between female teachers and male teachers' tendency that finds alternative assessment supporting learner-centered education (SAAM) (p <.05). This significant difference occurred at the small

effect size. Female teachers' tendency that finds alternative assessment supporting learner-centered education (SAAM) is higher than that of male teachers. A significant difference was determined between female teachers and male teachers' tendency to use alternative assessment methods ($p < .05$). This significant relationship occurred at a small effect size. Female teachers' tendency to use alternative assessment methods is higher than male teachers' tendency to use alternative assessment methods.

Teachers' tendency to use alternative assessment methods by year of seniority

Teachers' tendency levels to use alternative assessment methods according to their year of seniority were compared through One-Way ANOVA. The results are shown in Table 15.

The results of the analysis revealed that there is no significant difference in terms of teachers' tendency that finds alternative assessment methods useful (UAAM), tendency that finds alternative assessment methods supporting learner-centered education (SAAM), and tendency to use alternative assessment methods according to the seniority of the teachers ($p > .05$). However, a significant relationship was noted in terms of finding alternative assessment methods impractical (IAAM) according to the seniority of the teachers ($p < .05$). This significant difference is at the small effect size level. The teachers with 6-10 years of seniority, 16-20 years of seniority, and 21 years and above of seniority do not use alternative assessment methods as much as the teachers with 0-5 years of seniority. In this case, it can be interpreted that an increase in the year of seniority causes a decrease in the tendency to use alternative assessment methods.

Table 15: Teachers' tendency to use alternative assessment methods by year of seniority (One-Way ANOVA)

Factors	*Year of Seniority*	*N*	*Mean (S. Dev.)*	*F*	*df*	*p*	*ηp*	*Significant Difference*
Approach that finds alternative assessment methods useful (UAAM)	0-5	659	21.45(4.6)	0.257	4	0.905	0.001	No
	6-10	230	21.43(4.7)					
	11-15	79	21.70(4.1)					
	16-20	32	20.78(4.6)					
	21 and above	22	21.82(5.6)					
Approach that finds alternative assessment impractical (IAAM) (*)	0-5	659	8.68(3.8)	5.217	4	<0.0001	0.020	0-5<6-10
	6-10	230	9.71(4.2)					0-5<16-20
	11-15	79	9.49(3.2)					
	16-20	32	10.72(3.8)					0-5<21 and above
	21 and above	22	9.95(5)					
Approach that finds alternative assessment supporting for learner-centered education (SAAM)	0-5	659	15.28(3.3)	1.467	4	0.210	0.006	No
	6-10	230	14.86(3.4)					
	11-15	79	15.05(3.2)					
	16-20	32	15.78(2.7)					
	21 and above	22	16.14(2.7)					
Tendency to Use Alternative Assessment Methods (**)	0-5	659	52.05(9.1)	1.472	4	0.209	0.006	No
	6-10	230	50.58(9.3)					
	11-15	79	51.25(8.4)					
	16-20	32	49.84(7.4)					
	21 and above	22	52.00(9.1)					

(*) Calculated without reverse coding. High scores indicate a negative tendency/approach to use alternative assessment methods.

(**) After negative approach items were coded reversely, total score was obtained. High scores indicate a positive tendency/approach to use alternative assessment methods.

Discussion, Conclusion and Suggestions

This study aimed to develop a measurement tool (T-TUAS) to determine teachers' tendency to use alternative assessment methods and examined potential variations among teachers' use of these methods based on various sociodemographic variables in ……. The study revealed that teachers, in general, have a positive tendency to alternative assessment methods. This finding is in line with the research in the related literature (Ahmedi, 2019; Barrientos Hernán et al., 2023; Cadawas, 2024; Kansızoğlu et al., 2024; Kippers et al., 2019; Nasri et al., 2010; Shahbari & Abu-Alhija, 2018; Şahin & Öztürk, 2014; Yıldırım, 2023).

Another finding was that the most common three alternative assessment techniques that teachers use were *concept mapping*, *observation,* and *interview*. This may be due to the fact these techniques do not require much time to prepare, or they may feel more self-confident in preparing the tools for these techniques. Another reason might be the teachers had in-service training about the preparation and use of these tools. This finding is in line with the study conducted by Çalışkan and Kaşıkçı (2010). Çalışkan and Kaşıkçı (2010) stated that teachers feel safer to use alternative assessment methods if they have a previous in-service training. Similarly, Özdemir and Nakiboğlu (2023) found that the teachers who took courses or training in undergraduate teacher education were found to have high self-confidence in using alternative assessment tools. However, we should also note that teacher knowledge of alternative assessment alone is not enough. Other factors such as class size or number of students should be considered. For instance, classrooms in the state schools, especially in the eastern part of ……., are over-crowded (Gökçe et al., 2017). Therefore, the teachers may not spare enough time to use alternative assessment tools although they are aware of the fact that alternative assessment is advantageous. In addition, teachers' workload may impact the practical implementation of alternative assessment methods in the classroom.

Given the comparisons made through *gender* and *year of seniority* in terms of the level of teachers' tendency to use alternative assessment methods, significant differences were found. In terms of gender, although the significant difference between female and male teachers' tendency to use alternative assessment methods impractical (IAAM) was at the small effect size, the tendency level to use alternative assessment of male teachers who find the use alternative assessment methods impractical (IAAM) is higher than female teachers. In addition, a significant difference was noted between female teachers and male teachers in terms of finding alternative assessment supporting for learner-centered education (SAAM), which was at the small effect size. Female teachers who find alternative assessment supporting for learner-centered education (SAAM) was higher than that of male teachers. Similarly, there was a significant difference between female teachers' and male teachers' tendency to use alternative assessment methods ($p < .05$). Female teachers' tendency to use alternative assessment methods is higher than that of male teachers. This finding aligns with the study conducted by Kaya, Balay and Göçen (2012) in that female teachers use alternative assessment methods more than male teachers. However, Kuran and Kanatlı (2009) found that male teachers have a more positive attitude to use alternative assessment methods in their classrooms. In another study (Ak & Güvendi, 2010; Yusron et al., 2024), it was noted that there is no significant difference between gender groups in terms of using alternative assessment methods in the classroom.

On the other hand, the analysis by teachers' *year of seniority* revealed that there is no significant relationship in terms of teachers' tendency that finds alternative assessment methods useful (UAAM), tendency that finds alternative assessment methods supporting learner-centered education (SAAM), and tendency to use alternative assessment methods. However, a significant difference was found in terms of finding alternative assessment methods impractical (IAAM) according to the seniority of the teachers. The teachers with 6-10 years of seniority, 16-20 years of seniority, and 21 years and above of seniority do not use alternative assessment methods as much as the teachers with 0-5 years of seniority. In this respect, it can be concluded that an increase in the year of seniority causes a decrease in the tendency to use alternative assessment methods. This may be because novice teachers may have higher expectations about the school (perhaps driven by a desire to innovate, become a change agent, and cater to diverse learning styles) in the first years, but it gradually disappears as those teachers gain more experience and become accustomed to the existing educational system. However, this finding does not align with the study by Ak and Güvendi (2010). Ak and Güvendi (2010) found that the year of seniority does not have any impact on teachers' use of alternative assessment methods. Similarly, the results of the study carried out by Dilmaç and Dilmaç (2020) revealed that the year of seniority was not a determinant in teachers' attitudes toward using alternative assessment methods. On the contrary, Kaya, Balay, and Göçen (2012) noted in their study that teachers with 21 or more year of seniority use alternative assessment methods more in their classrooms compared to the teachers with lower years of seniority.

In conclusion, we believe that our study provides valuable insights into teachers' tendencies to use alternative assessment

methods, shedding light on the factors that influence their adoption in educational settings. The findings emphasize the need for professional development opportunities and institutional support to facilitate the integration of alternative assessment methods in learning environments. Therefore, it's essential for teachers, school administrators, and educational institutions to recognize the importance of alternative assessment and foster a culture that encourages continuous learning and adaptation to assessment for learning, regardless of seniority, which will pave the way for continuous monitoring and collecting evidence of learners' progress during the learning and teaching process. To do this, professional development opportunities such as in-service training must be provided for teachers to improve positive attitudes and thus increase teacher practice on alternative assessment methods.

For future research, further exploration is needed to explore the long-term impact of alternative assessments on learning and achievement, and overall quality in education. In addition, the nuanced reasons behind the variations in teachers' tendencies toward alternative assessment methods and investigating the challenges that teachers or practitioners face in the implementation of these methods could offer practical solutions for educators and educational policy makers in the educational institutions. Researchers can conduct studies that further examine innovative and alternative assessment strategies and methods that promote deeper learning for students. Additionally, a qualitative study can be conducted with the teachers for an in-depth analysis of the reasons for their choice to use alternative assessment techniques. Additionally, investigating the impact of external factors such as a centralized education system based on standardized central exams, class size, and teachers' workload on the practical implementation of these methods is essential.

Limitations

This study has the following limitations. First, although the study was carried out with a large number of participants (1022 teachers who work at primary, middle, or high schools), all data were collected in Şırnak, Turkey. Second, this measurement tool was developed to be used with all educators no matter what educational level they are teaching. However, we could not collect any data from the educators who work at universities. Despite these limitations, the current study presents significant results and some useful information regarding the use of alternative assessment methods by K-12 teachers.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics Statement (Including the Committee Approval Number)

This study was conducted with the approval of Şırnak University Scientific Research Ethics Committee (Date of Approval: 07/12/2023-No: 2023/82405). And, to implement the measurement tool at schools, the research application permission was obtained from Şırnak Governorship Provincial Directorate of National Education (Date: 20/12/2023-No: E-61543340-604.01.01-92552958).

References

Ahmad, S., Sultana, N., & Jamil, S. (2020). Behaviorism vs constructivism: A paradigm shift from traditional to alternative assessment techniques. *Journal of Applied Linguistics and Language Research*, *7*(2), 19-33. https://www.jallr.com/index.php/JALLR/article/view/1092/1261

Ahmedi, V. (2019). Teachers' attitudes and practices towards formative assessment in primary schools. *Journal of Social Studies Education Research*, *10*(3), 161-175. https://www.learntechlib.org/p/216460/

Ak, E., & Güvendi, M. (2010). Assessment of the degree to which primary school teachers use alternative assessment and evaluation methods. *Procedia-Social and Behavioral Sciences*, *2*(2), 5599-5604. https://doi.org/10.1016/j.sbspro.2010.03.913

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness of fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173. https://doi.org/10.1007/BF02294170

Austin, H., Dwyer, B., & Freebody, P. (2005). *Schooling the child: The making of students in classrooms*. Routledge.

Ayaz, M., & Gök, B. (2023). The effect of e-portfolio application on reflective thinking and learning motivation of primary school teacher candidates. *Current Psychology*, *42*(35), 31646-31662. https://doi.org/10.1007/s12144-022-04135-2

Ayyoub, A. A., Shamali, M., Salih, S., & Jabali, O. (2022, March). The Effect of Alternative Assessment in Scientific Thinking in Light of the Corona Pandemic Among Students of the Upper Basic Stage in Nablus Governorate. In *International Conference on Business and Technology* (pp. 210-218). Cham: Springer International Publishing.

Bentler P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Billington, R. (2003). *Living Philosophy: An introduction to moral thought*. New York: Routledge.

Brualdi, A. C. (1996). Multiple intelligence: Gardner's Theory: Washington DC: ERIC Clearing house on Assessment and Evaluation. https://files.eric.ed.gov/fulltext/ED410226.pdf

Buhagiar, M.A. (2007). Classroom assessment within the alternative assessment paradigm: revisiting the territory. *Curriculum Journal, 18*(1), 39–56. https://doi.org/10.1080/09585170701292174

Cadawas, R. M. (2024). Academic performances in social science through comprehensive learning portfolio. *International Journal of Education, Technology and Science, 4*(4), 2274–2284.

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6). https://doi.org/10.18637/jss.v048.i06

Corcoran, C. A., Dershimer, E. L., & Tichenor, M. S. (2004). A teacher's guide to alternative assessment: Taking the first steps. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *77*(5), 213-218. https://doi.org/10.3200/TCHS.77.5.213-218

Cowan, J., & Cherry, D. (2012). The learner's role in assessing higher level abilities. *Practitioner Research in Higher Education*, *6*(1), 12-22. http://insight.cumbria.ac.uk/id/eprint/1327/

Culbertson, L. D., & Wenfan, Y. (2003, April). *Alternative Assessment: Primary Grade Literacy Teachers' Attitudes and Practices*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. https://files.eric.ed.gov/fulltext/ED479794.pdf

Çalışkan, H., & Kaşıkçı, Y. (2010). The application of traditional and alternative assessment and evaluation tools by teachers in social studies. *Procedia-Social and Behavioral Sciences*, *2*(2), 4152-4156. https://doi.org/10.1016/j.sbspro.2010.03.656

Dilmaç, S., & Dilmaç, O. (2020). Visual Art Teachers' Determination of the Self-Sufficiency to Use Alternative Assessment Tools. *International Journal of Evaluation and Research in Education, 9*(2), 292-302. https://doi.org/10.11591/ijere.v9i2.20496

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(SUPPL. 1), 5-18). https://doi.org/10.1007/s11136-007-9198-0

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. California: SAGE Publications, Inc.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Research in Personality, 29*(2), 168-188. https://doi.org/10.1006/jrpe.1995.1010

Gibbs, G. (2019). How assessment frames student learning. In *Innovative Assessment in Higher Education* (pp. 22-35). Routledge.

Gipps, C. V., & Stobart, G. (2003). Alternative assessment. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 549–575). Dordrecht: Springer. https://doi.org/10.1007/978-94-010-0309-4_33.

Gökçe, N., Kaya, E., Aktaş, S. G., & Kantar, Y. M. (2017). Regional Research of Classroom Needs in Secondary Schools in Turkey-2013, *Journal of Enterprise Resource Planning Studies, 2017*(2017), 1-9. https://doi.org/10.5171/2017.661323

Gronlund, N.E. (1998). *Assessment of student achievement*. Boston: Allyn and Bacon.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis*. Harlow: Pearson Education Limited.

Hambleton, R K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publication.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological test: A progress report. *European Journal of Psychological Assessment, 10*(3), 229-244.

Herman, J. L. (1992). *A practical guide to alternative assessment*. Association for Supervision and Curriculum Development, 1250 N. Pitt Street, Alexandria, VA 22314.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*(1), 53-60.

Hu L. T., & Bentler P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling *A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement, 34*, 111-117. https://doi.org/10.1177/001316447403400115

Kansızoğlu, N., Kansızoğlu, H. B., & Karataş, F. Özgür. (2024). Turkish Language Teachers' Formative Assessment Competencies and Barriers for Formative Assessment: A Mixed-Methods Study. *Pegem Journal of Education and Instruction, 14*(4), 457–473. https://doi.org/10.47750/pegegog.14.04.44

Kaya, A., Balay, R., & Göçen, A. (2012). Öğretmenlerin alternatif ölçme ve değerlendirme tekniklerine ilişkin bilme, uygulama ve eğitim ihtiyacı düzeyleri. *International Journal of Human Sciences*, *9*(2), 1229-1259. https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/2272/982

Kemmis, S., & Edwards-Groves, C. (2018). Studying Education. In *Understanding Education* (pp. 1-30). Springer, Singapore.

Kippers, W. B., Wolterinck, C. H., Schildkamp, K., Poortman, C. L., & Visscher, A. J. (2018). Teachers' views on the use of assessment for learning and data-based decision making in classroom practice. *Teaching and Teacher Education*, *75*, 199-213. https://doi.org/10.1016/j.tate.2018.06.015

Kline, T. J. B. (2005). *Psychological testing, a practical approach to design and evaluation*. The USA: Sage.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An r package for assessing multivariate normality. *The R Journal, 6*(2):151-162. https://doi.org/10.32614/RJ-2014-031

Kuran, K., & Kanatlı, F. (2009). Alternatif ölçme değerlendirme teknikleri konusunda sınıf öğretmenlerinin görüşlerinin değerlendirilmesi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 6*(12), 209-234. https://dergipark.org.tr/en/download/article-file/183111

Liu, M., Kitto, K., & Shum, S. B. (2021). Combining factor analysis with writing analytics for the formative assessment of written reflection. *Computers in Human Behavior*, *120*, 106733. https://doi.org/10.1016/j.chb.2021.106733

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indices in confirmatory factor analysis: The effect of sam-

ple size. *Psychological Bulletin, 103*(3), 391-410. https://doi.org/10.1037/0033-2909.103.3.391

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Hillsdale, NJ: Erlbaum.

Mofolo, N. (2023). *Evaluation of the implementation and assessment of AfriMEDS physician competency framework in an undergraduate medical programme: a South African mixed-methods case study* (Doctoral dissertation, University of Bath) https://purehost.bath.ac.uk/ws/portalfiles/portal/275718426/N_Mofolo_DBA_Thesis_14_04_2023_final.pdf

Morgado, F. F., Meireles, J. F., Neves, C. M., Amaral, A. C., & Ferreira, M. E. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia*, *30*(1), 3–20. https://doi.org/10.1186/s41155-016-0057-1

Nasri, N., Roslan, S. N., Sekuan, M. I., Bakar, K. A., & Puteh, S. N. (2010). Teachers' perception on alternative assessment. *Procedia-Social and Behavioral Sciences*, *7*, 37-42. https://doi.org/10.1016/j.sbspro.2010.10.006

Nitko, A. J., & Brookhart, S.M. (1996). *Educational assessment of students*. Pearson Education Limited, Edinburgh Gate.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Orakçı, Ş. (2021). Teachers' reflection and level of reflective thinking on the different dimensions of their teaching practice. *International Journal of Modern Education Studies*, *5*(1), 118-139. http://dx.doi.org/10.51383/ijonmes.2021.88

Özdemir, M., & Nakiboğlu, C. (2023). Investigation of the Chemistry and Science Teachers' Self-Efficacy on Alternative Assessment and Evaluation in Terms of Some Variables. *Turkiye Kimya Dernegi Dergisi, 8*(2), 137-156. https://doi.org/10.37995/jotcsc.1354350

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Virginia: SAGE Publications, Inc.

Pifer, D. A. (2000). Getting in Trouble: The Meaning of School for "Problem" Students. *The Qualitative Report, 5*(1), 1-26. Retrieved from http://nsuworks.nova.edu/tqr/vol5/iss1/7

Randall, B. & Good, J.W. (2004). Educare and educere: Is a balance possible in the educational system? *Educational Forum, 68*(2), 161-168.

Schuwirth, L. W., & Van der Vleuten, C. P. (2011). Programmatic assessment: from assessment of learning to assessment for learning. *Medical Teacher*, *33*(6), 478-485. https://doi.org/10.3109/0142159X.2011.565828

Shahbari, J.A., &Abu-Alhija, F.N. (2018). Does Training in Alternative Assessment Matter? The Case of Prospective and Practicing Mathematics Teachers' Attitudes Toward Alternative Assessment and Their Beliefs About the Nature of Mathematics. *Int J of Sci and Math Educ, 16*, 1315–1335. https://doi.org/10.1007/s10763-017-9830-6

Stacey, K., Price, B., Steinle, V., Chick, H., & Gvozdenko, E. (2009, September). SMART assessment for learning. In *Conference of the International Society for Design and Development in Education, Cairns, Australia*.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Pearson Education.

Tawil, S., & Locatelli, R. (2015). *Rethinking education: towards a global common good*. Paris: UNESCO.

Thissen, D. (1991). *Multilog User's Guide*. Chicago: Scientific Software

Vieira A. L. (2011). *Interactive LISREL in practice, getting started with a SIMPLIS Approach*. London: Springer. https://doi.org/10.1007/978-3-642-18044-6

Warner, R. M. (2013). *Applied statistics, from bivariate through multivariate techniques*. California: SAGE Publications, Inc.

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 228-260. https://doi.org/10.1080/0969594X.2021.1884042

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3),187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yıldırım, Ş. S. (2023). A critical assessment of the EFL curriculum and the course book for fourth grade students in light of the CERF. *International Journal of Education, Technology and Science, 3*(3), 1078–1111.

Yurdabakan, İ. (2011). The view of constructivist theory on assessment: Alternative assessment methods in education. *Ankara University Journal of Faculty of Educational Sciences (JFES)*, *44*(1), 51-78. https://doi.org/10.1501/Egifak_0000001215

Yusron, E., Istiyono, E., Hidayati, K., Retnawati, H., & Hassan, A. (2024). Teacher's Perspective on Diagnostic Assessment Process in the Midst of Covid-19. *Pegem Journal of Education and Instruction, 14*(4), 400–409. https://doi.org/10.47750/pegegog.14.04.38

Zhao, Y. (2008). *Approaches for addressing the fit of item response theory models to educational test data*. (Doctoral Dissertation). University of Massachusetts Amherst.