International Journal of Assessment Tools in Education
2025, Vol. 12, No. 3, 806–824
https://doi.org/10.21449/ijate.1606780
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Development and validation of digital assessment literacy self-efficacy scale

Ruhan Karadağ Yılmaz[1], **İlhan Koyuncu**[2*]

[1]Selçuk University, Faculty of Education, Department of Elementary Education, Konya, Türkiye
[2]Adıyaman University, Faculty of Education, Department of Educational Sciences, Adıyaman, Türkiye

ARTICLE HISTORY
Received: Dec. 24, 2024
Accepted: June 2, 2025

Keywords:
Assessment literacy,
Digital literacy,
Digital assessment,
Scale development,
Self-efficacy.

Abstract: This study aimed to develop a scale to measure teachers' digital assessment literacy self-efficacy. Teachers were selected from all regions of Turkey and different branches at primary, secondary, and high school levels to enhance generalization and diversity. The data was collected from 314 teachers for the exploratory factor analysis and 296 for the confirmatory factor analysis. Various evidence was obtained regarding the validity and reliability of the Digital Assessment Literacy Self-Efficacy Scale (DALSS). The exploratory factor analysis results revealed that DALSS was found to have a 3-factor structure with 24 items. The dimensions were named as "Using appropriate tool and assessment type", "Feedback and developing appropriate tools for the objectives and levels," and "Preparing exams and evaluating the usability of the scores." Confirmatory factor analysis (CFA) results showed that the hypothetical structure of the scale fit the data. Second-order CFA also confirmed the 3-factor structure of digital assessment literacy self-efficacy. Convergent and divergent validity results proved evidence for construct validity. Analyses based on known-group validity revealed that the DALSS is a discriminative instrument. The internal consistency reliability values of DALSS and its factors were found to vary between .96 and .98. The findings for practice revealed that although DALSS scores did not differentiate according to gender, professional experience and teaching level, they significantly differed in terms of teaching fields.

1. INTRODUCTION

Teachers' assessment knowledge and competence is one of the most important factors contributing to the effectiveness of education (Nimehchisalem & Bhatti, 2019). Teachers are expected to use various assessment tools, integrate multiple assessment forms into the teaching process to measure students' progress, and develop and maintain a sound understanding of assessment practices and theories (DeLuca & Klinger, 2010). In other words, teachers need to possess specialized assessment knowledge and skills (Edwards, 2016; Rogier, 2014), to be skilled in many aspects of assessment literacy to achieve the desired goals of assessment (Al-Bahlani, 2019), and to have the assessment literacy to select the most appropriate assessment strategy for students' success (Nyagi & Rajendran, 2020).

*CONTACT: İlhan KOYUNCU ilhankync@gmail.com Adıyaman University, Faculty of Education, Department of Educational Sciences, Adıyaman, Türkiye

e-ISSN: 2148-7456

Assessment literacy has become a major focus as a core professional requirement in many education systems (DeLuca *et al.*, 2016a, 2016b). Due to the crucial role of assessment practices within the education system, teachers need to be assessment literate (Edwards, 2016; Shams *et al.*, 2018). Today, teachers need to be digitally assessment literate to evaluate students' performance using modern digital technologies (Eyal, 2012; Rezai *et al.*, 2021).

Receiving feedback on teachers' competencies in digital assessment literacy will help them develop their digital assessment literacy. However, the review of the extant literature indicates an absence of any scale that aims to measure teachers' digital assessment literacy self-efficacy. Therefore, this study aimed to develop a scale to determine teachers' digital assessment literacy self-efficacy.

1.1. Assessment Literacy

At the present time, no universally accepted definition of assessment literacy exists (Walters, 2010). The term was first coined by Stiggins (1991) and refers to the knowledge and skills teachers need to conduct assessments effectively (Fulcher, 2012). Popham (2011) defines assessment literacy as an individual's understanding of basic assessment concepts and procedures that are likely to influence educational decisions, while Brookhart (2011) and Volante and Fazio (2007) describe it as the knowledge and skills required for teachers to measure and assess student learning. Competencies of assessment-literate individuals include understanding basic assessment concepts and procedures (Popham, 2011), designing assessment instruments, collecting relevant data using appropriate tools, analyzing and interpreting the collected data, and utilizing assessment results meaningfully (Yamtim & Wong-Wanich, 2014). Over the last two decades, teacher assessment literacy has emerged as a significant area of research, with the development of assessment literacy becoming a priority for many universities worldwide (Chan & Luo, 2020). However, an important facet that has not been extensively explored in assessment literacy research is digital assessment literacy. Today, teachers require assessment literacy that is adapted to the digital environment and aligns with 21st-century pedagogical approaches (Eyal, 2012). Consequently, teachers must possess the necessary digital assessment literacy to effectively evaluate students' performance using modern digital technologies, allowing them to be competent assessors in the digital age (Eyal, 2012; Rezai *et al.*, 2021).

1.2. Digital Assessment Literacy and Self-Efficacy

In recent years, the implementation of online assessments, computer-based testing, and learning management systems has made digital assessment competencies and practices increasingly important (Al-Bahlani and Ecke, 2023). This evolution has led to the emergence of the concept of digital assessment literacy. Digital assessment literacy generally refers to the knowledge and skills necessary for effectively utilizing digital tools in assessment processes. A digital assessment literate teacher should possess the skills and abilities to use assessment data from a digital database for the planning of teaching-learning-assessment processes. This includes effectively employing digital tools at all stages of the assessment process, managing formative and summative assessment scores based on a digital database, and interpreting the results effectively (Eyal, 2012). There is a pressing need to investigate how well-prepared teachers are to administer assessments in a digital environment (Al-Bahlani & Ecke, 2023). In other words, it is crucial to examine teachers' self-efficacy regarding digital assessment literacy. Digital assessment literacy self-efficacy refers to educators' knowledge, skills, and attitudes necessary for effectively using digital technologies in educational assessments. This includes the ability to understand and manage digital assessment processes as well as the capacity to utilize digital tools for assessment purposes in education.

Teachers' digital assessment literacy self-efficacy is crucial for the success of modern education systems and student achievement. A low level of digital assessment literacy self-efficacy may indicate difficulties in adapting to modern technologies and digital tools, highlighting potential

gaps in professional development and support. This situation may also reflect inadequacies in education policies and strategies that fail to support the effective use of digital tools. Therefore, enhancing teachers' digital assessment literacy and strengthening their self-efficacy in this area are critical to improving educational quality and creating a more effective learning environment. Measuring teachers' digital assessment literacy self-efficacy is essential in this context to identify areas for improvement and to guide targeted professional development efforts.

When the studies on assessment literacy are examined, it is noteworthy that studies on this subject have gained importance and become widespread in recent years. Although numerous studies aim to measure the assessment literacy of teachers and pre-service teachers, there remains a notable gap in the development of scales specifically designed for measuring assessment literacy and digital assessment literacy. While some scales have been developed to assess various aspects of assessment literacy, they often fall short of addressing contemporary needs. Edwards (2016) developed a rubric to monitor the development of pre-service teachers' assessment literacy. DeLuca *et al.* [2016a] developed an assessment literacy tool for teachers and researchers to use. Most studies on teacher assessment knowledge have used the Teacher Assessment Literacy Questionnaire (TALQ) developed by Plake and Impara (1992) based on the 1990 standards or a modified version of it (Al-Bahlani, 2019). However, DeLuca *et al.* (2016a) noted that many of these tools do not adequately reflect recent changes in the assessment environment. Given the significant advancements in classroom assessment over the past 20 years, there is a clear need for tools that incorporate contemporary assessment practices and address multiple dimensions of assessment literacy. Similarly, other studies Al-Bahlani, 2019; Brookhart, 2011) have emphasized that existing tools are outdated and fail to meet current assessment demands. Al-Bahlani (2019) stated that digital assessment competence is rarely discussed when measuring teacher assessment literacy, so there is an undeniable need for research to measure digital assessment literacy.

As the effects of the digital age on education rapidly increase, measuring and developing teachers' competencies in their ability to make effective assessments with digital tools and platforms is becoming increasingly important. However, the absence of a scale development study on digital assessment literacy self-efficacy perception in literature creates a significant gap and deficiency. Without a scale that accurately measures teachers' self-efficacy perceptions on digital assessment, it is not possible to effectively design development studies and training programs in this area. A scale development study on teachers' digital assessment literacy self-efficacy perception is a critical step to increase the effectiveness of digital transformation in education and support teachers' professional development. This scale will contribute to determining teachers' self-efficacy in digital assessment literacy and manage teaching processes more effectively through the results obtained.

1.3. Purpose of the Study

This study aims to develop a scale with high validity and reliability that can reveal teachers' self-efficacy for digital assessment literacy. The Digital Assessment Literacy Self-Efficacy Scale (DALSS) can inform teachers about their competencies in assessment and their training needs in this regard. In addition, the results obtained through the scale can contribute to the realization of studies (projects, seminars, etc.) aiming to improve teachers' assessment literacy. The results of this study will also be an important resource for future researchers in determining teachers' digital assessment literacy.

2. METHOD

This section included information about the research model, the participants, the stages of the scale development process, and data analysis.

2.1. Research Model

The present study was a survey research since a scale for digital assessment literacy self-efficacy levels was developed and validated. Survey research, a quantitative research technique, is a type of research that aims to describe some aspects and characteristics of a representative sample selected from the target population (Fraenkel *et al.*, 2011).

2.2. Participants

The sample of the study consisted of teachers from different teaching fields at primary, secondary, and high school levels in the 2022-2023 academic year across Turkey. In sample selection, convenient sampling, which is a type of purposeful sampling method, was preferred. For the exploratory factor analysis (EFA), a sample size of 300 participants is suggested to obtain acceptable results (Comrey & Lee, 1992; Tabachnick & Fidell, 2012). Therefore, data were collected from 314 participants (Sample 1) for EFA. In addition, although the minimum required sample size for confirmatory factor analysis (CFA) depends on different conditions such as the number of items and factors, and the parameter estimation method, in general, a sample size of 300 or more is acceptable. In addition, it is suggested to perform EFA and CFA on different samples (Worthington & Whittaker, 2006). Hence, after obtaining the structure of the scale by using EFA, the scale was applied to a different sample of 296 (Sample 2) participants for CFA. Descriptive statistics of the participants are given in Table 1.

Table 1. *Descriptive statistics of the participants.*

		Sample 1		Sample 2	
		N	*%*	*N*	*%*
Gender	Female	173	55.1	166	56.1
	Male	141	44.9	130	43.9
	Total	314	100.0	296	100.0
Teaching field	Primary	142	45.2	122	41.2
	Turkish and social	47	15.0	55	18.6
	Science and mathematics	66	21.0	64	21.6
	Educational sciences	14	4.5	12	4.1
	Special education	5	1.6	5	1.7
	Foreign languages	13	4.1	9	3.0
	Other (vocational, art, philosophy, informatics etc.)	27	8.6	29	9.8
	Total	314	100.0	296	100.0
Professional experience (years)	0-5	139	44.3	122	41.2
	6-10	54	17.2	51	17.2
	11-15	38	12.1	47	15.9
	16-20	34	10.8	31	10.5
	21 and above	49	15.6	45	15.2
	Total	314	100.0	296	100.0
Training on digital assessment	Yes	111	35.4	109	36.8
	No	203	64.6	187	63.2
	Total	314	100.0	296	100.0
Teaching level	Primary	138	43.9	113	38.2
	Middle school	78	24.8	91	30.7
	High school	98	31.2	92	31.1
	Total	314	100.0	296	100.0
Age		*M*	*SD*	*M*	*SD*
		32.88	9.62	33.50	8.83

According to Table 1, the distribution of participants according to their gender, teaching field, professional experience, training on digital assessment, and teaching level was similar across the two samples. While the number of male and female participants was close to each other, the participants mostly consisted of female teachers. The participants mostly consisted of primary school, science and mathematics, Turkish and social science teachers. Most of the participants had less than 10 years of professional experience and did not receive any training on digital assessment. The participants were from all teaching levels of the formal education process. The ages of the participants in Sample 1 ranged between 22 and 60, with a mean of 32.88 (*SD* = 9.62), and for Sample 2, it ranged between 22 and 59, with a mean of 33.50 (*SD* = 8.83).

2.3. Procedures

In the scale development process, the item generation, theoretical analysis, and psychometric analysis stages suggested by Morgado *et al.* (2018) were followed. The psychometric analysis stage was included in the data analysis subheading.

2.3.1. *Item generation and determination of the scale framework*

Item generation is the stage where theoretical support is provided for the item pool by conducting a comprehensive literature review (Boateng *et al.*, 2018). At this stage, the theoretical framework and scope of DALSS and the competencies that teachers with digital assessment literacy should possess are yet to be determined. In this process, the dimensions that should be included in the scale were emphasized by the researchers, and the purpose and scope of the scale were decided. While the digital assessment literacy scale was being prepared, the purpose and scope of the scale were determined in line with the existing assessment literacy scales (Edwards, 2016; DeLuca *et al.*, 2016a; Mertler, 2004; Mertler & Campbell, 2005; Plake *et al.*, 1993; Zhang & Burry-Stock, 2003) and teacher standards for educational assessment of students (Stiggins, 1999; AFT *et al.*, 1990; MoNE, 2017). As a result of the literature review, a draft form of the DALSS scale, consisting of 34 different items aiming to measure teachers' digital assessment literacy was generated by the researchers. The items were developed originally in the Turkish language (Appendix A). A translation of the items to the English was given in Table 3.

2.3.2. *Theoretical analysis phase*

At this stage, the researcher assessed the content validity of the new scale and ensured that the initial item pool reflected the desired construct. Content validity was mainly assessed through evaluations by experts and target participant groups. To ensure content validity, expert opinions were obtained about the items created (Boateng *et al.*, 2018). For the theoretical analysis phase of the DALSS, the draft form consisting of 34 items was asked for the opinions of 7 different experts working as lecturers and faculty members in 7 different state universities; 3 of them were experts in measurement and evaluation, 1 in educational programs and teaching, 2 in computer and instructional technologies, and 1 in language. Then, Lawshe's (1975) content validity index (CVI) was calculated to examine whether the agreement between the expert opinions was statistically significant. In the present study, the CVI value was calculated as 1.00, and since this value was greater than the minimum required CVI value of .99 at .05 significance level (one-tailed), the content validity of the scale was found to be statistically significant.

In line with the opinions and suggestions of the experts regarding the statements in the first form of the draft scale, similar and/or synonymous items were combined, items that were not deemed necessary for the scale were removed, and a draft scale consisting of 31 items was formed (Appendix B). In addition, all items in the draft scale form were submitted by the researchers to the opinion of a language expert in terms of meaning and grammar, and necessary corrections were made in line with these suggestions. Finally, the scale items were scored on a 5-point Likert scale as *Completely Disagree* (1), *Disagree* (2), *Somewhat Agree* (3), *Agree* (4) and *Completely Agree* (5).

Ethical approval was obtained from institutional ethics board. The participants were informed about the scale, their consent was obtained, and voluntary participation was encouraged. A pre-application was conducted to see the applicability and comprehensibility of the draft form of the DALSS, to obtain the opinions and suggestions of the individuals participating in the application, and to identify problems that may occur during the pilot application in advance. Erkuş (2012) recommends a pre-application on a small group to see the applicability of the scale. For this purpose, after obtaining the necessary permissions, a pre-test was conducted with a group of 30 participants. As a result of the preliminary test, it was seen that the scale items were applicable in terms of comprehensibility and appropriateness to the level and format. After the final version of the scale form was checked, Google Forms web link of the scale was delivered to the participants via their colleagues from all geographical regions of Turkey to send teachers teaching at different teaching levels, having different professional experiences and teaching fields. Data collection process was managed from the analytics of Google Forms to ensure diversity of the participants.

2.4. Data Analysis

At this stage, it was aimed to collect evidence for the construct validity and reliability of the scale. For this purpose, the mean, standard deviation, skewness, and kurtosis values of the scale items and item-rest correlations were examined. In addition, it was tested whether the lower 27% and upper 27% groups differed significantly for each item on the total scale scores. Then, the exploratory and confirmatory factor analyses were conducted to examine construct validity. Before proceeding to factor analysis, the assumptions of the analyses were examined, and the results were included in the factor analysis subheading. Convergent and divergent validity of the scale were assessed by using Fornell and Larcker (1981) criterion. Cronbach's α and McDonald's ω were evaluated for internal consistency. In order to obtain more evidence for construct validity, validity evidence based on group differences (known-group validity) proposed by Cronbach and Meehl (1955) was examined. For this purpose, the differentiation of DALSS total scores in terms of whether teachers received training on digital assessment and evaluation was examined. The details about data analysis were provided in results section before each analysis. Analyses were conducted in Microsoft Excel 2007, Jamovi 2.5.6 (The jamovi project, 2024) and IBM SPSS Statistics 23.

3. RESULTS

In this section, the evidence obtained for the validity and reliability of the scale was presented under separate subheadings.

3.1. Item Analysis

In the preliminary analyses, the item means, standard deviations, skewness and kurtosis values and item-total correlations were calculated. Item statistics were given in Table 2. According to Table 2, the means of the scale items ranged between 3.10 and 4.34, and the standard deviations ranged between 0.838 and 1.234. Skewness and kurtosis values ranged between -2 and +2. Item-rest correlations ranged between .563 and .839. These values indicate that all items have sufficient levels of discrimination and have moderate or higher relationships with the scale total scores. The discrimination of the items in terms of total scores for the lower group and upper group each constituting 27% of the total group was also examined. Since the data deviated significantly from normal distribution according to histogram, skewness and kurtosis values (out of -1 and +1 bounds), and normality test ($p < .001$), Mann Whitney U test used for comparing lower and upper groups, and Spearman's rho correlation coefficient for evaluating item-total correlations. The results showed that item-total correlations ranged between .709 and .945 and were statistically significant ($p < .001$).

Table 2. *Descriptive statistics of the items.*

	M	*SD*	Skewness	*SE*	Kurtosis	*SE*	Item-rest correlation
item1	4.340	0.838	-1.343	0.097	1.901	0.195	0.563
item2	3.700	1.111	-0.570	0.097	-0.379	0.195	0.733
item3	3.670	1.106	-0.602	0.097	-0.257	0.195	0.775
item4	3.720	1.115	-0.720	0.097	-0.118	0.195	0.776
item5	3.810	1.079	-0.792	0.097	0.093	0.195	0.772
item6	3.580	1.097	-0.542	0.097	-0.269	0.195	0.771
item7	3.450	1.130	-0.457	0.097	-0.434	0.195	0.787
item8	3.720	1.103	-0.733	0.098	-0.109	0.195	0.818
item9	3.800	1.097	-0.736	0.097	-0.159	0.195	0.805
item10	3.570	1.178	-0.560	0.097	-0.532	0.195	0.698
item11	3.660	1.137	-0.545	0.097	-0.503	0.195	0.740
item12	3.590	1.234	-0.523	0.097	-0.721	0.195	0.748
item13	3.760	1.152	-0.714	0.098	-0.322	0.195	0.770
item14	3.720	1.123	-0.742	0.097	-0.141	0.195	0.830
item15	3.300	1.215	-0.298	0.097	-0.793	0.195	0.774
item16	3.100	1.199	-0.071	0.097	-0.832	0.195	0.718
item17	3.380	1.167	-0.366	0.097	-0.655	0.195	0.823
item18	3.450	1.143	-0.424	0.097	-0.496	0.195	0.834
item19	3.590	1.106	-0.546	0.097	-0.282	0.195	0.821
item20	3.690	1.144	-0.677	0.097	-0.271	0.195	0.838
item21	4.000	1.033	-1.018	0.097	0.593	0.195	0.784
item22	3.860	1.133	-0.910	0.097	0.076	0.195	0.802
item23	3.500	1.130	-0.424	0.097	-0.510	0.195	0.814
item24	3.400	1.163	-0.360	0.097	-0.638	0.195	0.799
item25	3.780	1.084	-0.738	0.098	-0.045	0.195	0.831
item26	3.660	1.116	-0.677	0.097	-0.151	0.195	0.839
item27	3.590	1.132	-0.582	0.097	-0.369	0.195	0.834
item28	3.410	1.200	-0.442	0.097	-0.595	0.195	0.823
item29	3.700	1.109	-0.740	0.097	-0.087	0.195	0.790
item30	3.770	1.084	-0.779	0.097	0.039	0.195	0.831
item31	4.220	0.982	-1.319	0.097	1.331	0.195	0.605

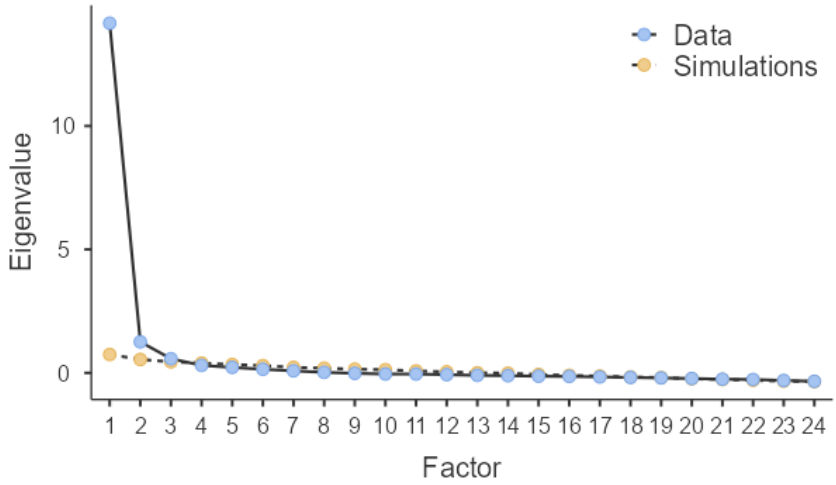
3.2. Exploratory Factor Analysis

Before conducting the exploratory factor analysis (EFA), Bartlett's test of sphericity and Keiser-Meier-Olkin (KMO) sampling suitability measures were examined. Tabachnick and Fidell (2012) suggested that a KMO value greater than .60 and a significant Bartlett's test of sphericity are sufficient for the exploratory factor analysis. Accordingly, Bartlett's test of sphericity ($\chi^2(465) = 9521$, $p < .001$) and KMO measure (.967) showed that the data were suitable for EFA. In addition, tolerance (> .01), variance inflation factor (VIF < 10), condition index (< 30) values showed that there was no multicollinearity. Mardia's test showed that the multivariate normality assumption was also met. Moreover, Mahalanobis distance values were calculated to determine multivariate outliers. It was found that there were not any significant outliers that affect EFA results. As factor extraction method principal axis factoring was utilized in combination with an oblimin rotation. Parallel analysis technique was used to decide on the number of factors. The initial results showed that the scale consisted of 4 factors (% of explained variance: Factor 1= 25.70, Factor 2=20.98, Factor 3=17.90, Factor 4=3.20, and Total=67.80).

However, when the factor loading values of the items were analyzed, item 20 seemed to have similar factor loadings for factors 1 and 3. Hence, this item was removed from the analysis. After deleting this item, the same situation was observed for item 12, and it was also removed. According to factor loadings, only items 21 and 31 had factor loadings for factor 4, and those values were lower than that of factor 1. Hence, the scale was considered to have 3 significant factors. Then, the number of factors was fixed to 3 and the analysis was repeated. After this adjustment, the factor loading value of item 31 declined below the cutoff value of .30. According to Tabachnick and Fidell (2012) items with factor loadings of .30 need to be removed from the analysis. After deleting items 31, item 21 was observed to have similar factor loadings for factors 2 and 3. When item 21 was removed, a similar situation was also observed for item 9. After removing item 9, the factor loading value of item 13 declined below the cutoff value of .30. Finally, item 13 was deleted from the analysis, and a scale consisting of 25 items and 3 factors was obtained. Before deleting all these items, their relevance for the scale was also assessed, and it was observed that there was no contraction in the sense of the scale.

Then, the factors were named as "Using appropriate tool and assessment type", "Feedback and developing appropriate tools for the objectives and levels", and "Preparing exams and evaluating the usability of the scores". Next, it was observed that although item 30 had factor loading for factor 1, it was not compatible with the name of the factor, and the objective of this item was also measured with items 14, 15, and 16. Hence, item 30 was also deleted, and the analysis was repeated. The final total explained value for the scale, consisting of 24 items with 3 factors, was 68%. The scale items were given in Table 3 and Appendix A. Total explained variances were 25.3% for factor 1, 21.8% for factor 2, and 20.9% for factor 3. The scree plot given in Figure 1 also confirms this result.

Figure 1. *Scree plot.*



According to the scree plot test (Cattell, 1978), the break point is determined by drawing lines from the two endpoints of the curve (Yong & Pearce, 2013). Although the scree plot showed that there was one factor that dominated the scale structure, parallel analysis indicated a 3-factor structure. Considering these results, factor loadings in Table 3 were obtained.

Tabachnick and Fidell (2012) recommended a minimum acceptable value of .30 for factor loadings. When Table 3 was examined, it was seen that the factor loading values varied between .321 and .881 and were at a sufficient level. The correlations between total scores and Factor 1 (r = .901), Factor 2 (r = .910), and Factor 3 (r = .926) were found to be very high and statistically significant (p < .001). Similarly, the correlation between Factor 1*Factor 2 (r = .732), Factor 1*Factor 3 (r = .744), and Factor 2*Factor 3 (r = .808) was found to be high and statistically significant (p < .001).

Table 3. *Scale items and their EFA factor loadings.*

	Items	λ	Uniqueness
Using appropriate tool and assessment	1. I can use information and communication technologies in the assessment process.	0.577	0.660
	2. I can use digital assessment tools to assess students.	0.573	0.394
	3. I can make diagnostic assessment practices in online environments.	0.721	0.324
	4. I can conduct formative assessment practices in online environments.	0.854	0.241
	5. I can make summative assessment practices in online environments.	0.881	0.205
	6. I can make criterion-referenced assessment practices in online environments.	0.649	0.335
	7. I can make norm-referenced assessment practices in online environments.	0.575	0.292
	8. I can involve students in assessment processes with digital assessment tools.	0.521	0.334
	10. I can include learning analytics from online learning systems in the assessment process.	0.321	0.549
	11. I follow new digital technologies related to the assessment process.	0.429	0.446
Feedback and developing	14. I can prepare assessment tools appropriate to the cognitive objectives of the course in digital environments.	0.596	0.307
	15. I can design digital assessment tools suitable for measuring affective skills.	0.768	0.270
	16. I can develop measurement tools for psychomotor objectives in digital environments.	0.694	0.371
	17. I can prepare digital assessment tools for different performance levels.	0.848	0.182
	18. I can prepare assessment tools suitable for students' developmental characteristics in digital environments.	0.778	0.192
	19. I can give feedback to all stakeholders according to the measurement and evaluation results I obtain in the digital environment.	0.608	0.353
Preparing exams and evaluating the	22. I can organize measurement and assessment activities in digital environments.	0.568	0.338
	23. I can determine the validity of the measurement results obtained in digital environment.	0.676	0.340
	24. I can determine the reliability of the measurement results obtained in the digital environment.	0.868	0.281
	25. I can use appropriate scoring methods for different assessment tools in the online environment.	0.802	0.238
	26. I can evaluate the findings from different assessment tools together in a digital environment.	0.772	0.220
	27. I can compare the advantages and limitations of online assessment tools.	0.812	0.222
	28. I can develop assessment tools appropriate to the structure of digital learning environments (synchronous or asynchronous).	0.652	0.250
	29. I can evaluate the usability of digital assessment tools.	0.574	0.345

Note. 'Principal axis factoring' extraction method was used in combination with a 'oblimin' rotation. Sum of square loadings: 5.01, % of variance: 68.0.

3.3. Confirmatory Factor Analysis

The assumptions of the confirmatory factor analysis (CFA) were checked before the analysis. Accordingly, it is recommended to have a sample of 300 or more (Worthington & Whittaker, 2006), to identify missing and extreme values, to examine univariate and multivariate normality, and to check singularity and multicollinearity (Koyuncu & Kılıç, 2019; Ullman, 2012). First of all, it can be said that a sample of 296 participants was almost appropriate for CFA. In addition, the sample size is also more than 10 times the number of items (24 items). Secondly, the data file was examined in detail, and no missing or erroneous data were found. There were no significant univariate and multivariate outliers according to Boxplots and Mahalanobis distances with their significance (p) values. Thirdly, although the skewness and

kurtosis values of each item for univariate normality were in the range of - 3 to +3, the histogram graphs showed that there were deviations from normality. On the other hand, Mardia's test showed that there was no deviation from multivariate normality. Finally, tolerance, variance inflation factor (VIF), and pairwise correlation values between items showed that there were no singularity and multicollinearity problems. After the assumptions of the confirmatory analysis were checked, the analyses were conducted with full information maximum likelihood (FIML) estimation method and factor variances were constrained to 1. FIML estimation method uses all information in the data and provides consistent and efficient estimates (Yuan, 2009).

To compare one-factor solution with three-factor solution, a CFA with one-factor were performed. The results revealed that that $\chi^2(252) = 1770$, $p < .001$, $\chi^2/df = 7.080$, Comparative Fit Index (CFI) = .791, Tucker–Lewis Index (TLI) = .771, Standardized-Root Mean Square Residual (SRMR) = .065 and Root Mean Square Error Approximation (RMSEA) = .139 (%90 CI with lower .132 and upper .145). For the fit indices, the χ^2/df value should be less than 5 (Anderson & Gerbing, 1984), and the SRMR value should be close to or less than 0.08 (Hu & Bentler, 1999). In addition, an RMSEA value is less than 0.10 (Browne & Cudeck, 1993), and CFI and TLI values between .90 and .95 indicate acceptable fit (Hu & Bentler, 1999). It was observed that one-factor solution did not fit to the data. Then, a CFA with three-factor solution was performed. It was found that $\chi^2(249) = 825$, $p < .001$, $\chi^2/df = 3.313$, CFI = .933, TLI = .926, SRMR = .038 and RMSEA = .088 (%90 CI with lower .077 and upper .097). The results showed that all fit values met the criteria given above. The three-factor solution provided a significantly better fit than the unidimensional alternative, thereby supporting the multidimensional nature of the construct. Factor loadings of the CFA model were given in Table 4.

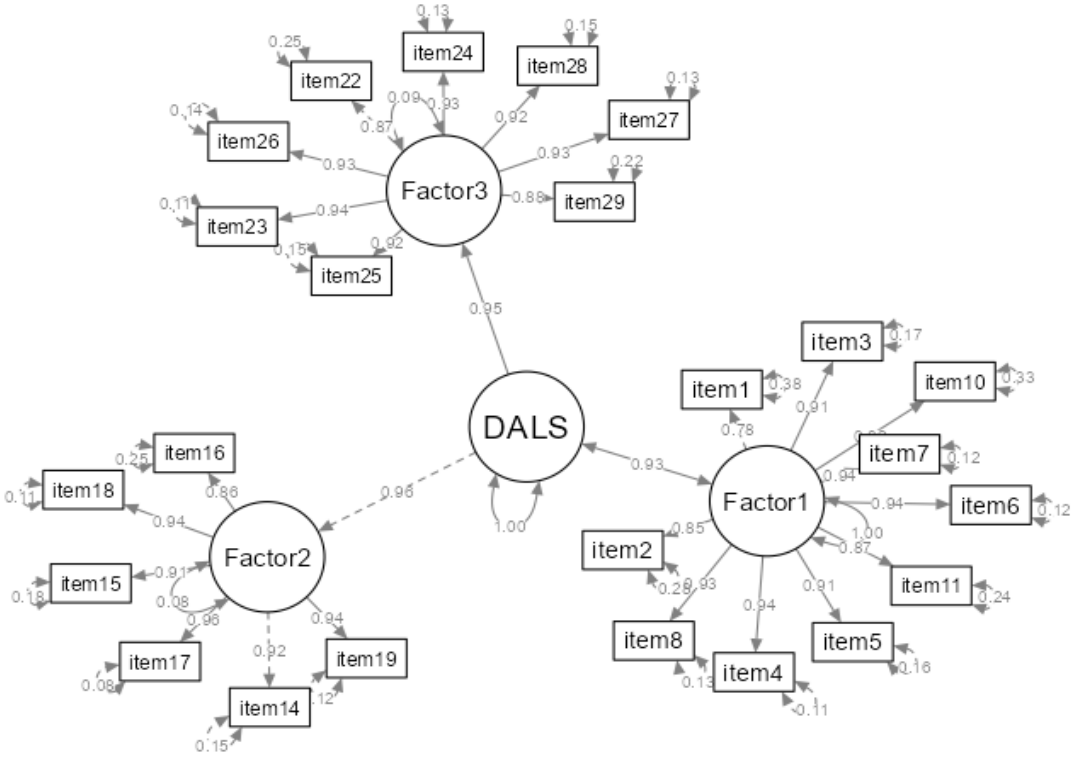
Table 4. *CFA factor loadings.*

Factor	Indicator	Estimate	*SE*	*z*	*p*	Stand. Estimate
Factor 1	item1	0.576	0.0438	13.2	< .001	0.678
	item2	0.880	0.0536	16.4	< .001	0.796
	item3	0.994	0.0517	19.2	< .001	0.880
	item4	1.022	0.0507	20.2	< .001	0.906
	item5	0.936	0.0489	19.1	< .001	0.878
	item6	0.996	0.0496	20.1	< .001	0.904
	item7	1.013	0.0508	19.9	< .001	0.900
	item8	1.003	0.0509	19.7	< .001	0.894
	item10	0.870	0.0586	14.8	< .001	0.742
	item11	0.920	0.0543	16.9	< .001	0.813
Factor 2	item14	0.967	0.0529	18.3	< .001	0.855
	item15	1.051	0.0543	19.4	< .001	0.885
	item16	0.995	0.0565	17.6	< .001	0.834
	item17	1.092	0.0515	21.2	< .001	0.933
	item18	1.056	0.0507	20.8	< .001	0.923
	item19	0.955	0.0509	18.7	< .001	0.869
Factor 3	item22	0.985	0.0557	17.7	< .001	0.835
	item23	1.063	0.0533	19.9	< .001	0.900
	item24	1.071	0.0538	19.9	< .001	0.899
	item25	0.960	0.0498	19.3	< .001	0.882
	item26	0.991	0.0493	20.1	< .001	0.904
	item27	1.042	0.0520	20.0	< .001	0.903
	item28	1.044	0.0560	18.6	< .001	0.864
	item29	0.960	0.0533	18.0	< .001	0.846

According to Table 4, factor loadings ranged between .678 and .933, and all items made a significant contribution to the model. When the CFA results were evaluated together, it was seen that the fit of the hypothetically tested model was acceptable, and the construct validity of

the scale was at a sufficient level. On the other hand, in order to understand whether 3-factor model represents digital assessment literacy self-efficacy as a unique construct, a second-order CFA was carried out. The diagonally weighted least squares method was used as the estimation technique. This method provides more accurate results in the case of ordinal data, and it is robust to deviations from normality and variable type (Mindrila, 2010). The analysis was carried out with lavaan (Rosseel, 2019), which is an R software (R Core Team, 2023) package by using Jamovi software. The results revealed that the proposed model had perfect fit to the data according to the fit values of $\chi^2 = 523$, $df = 249$, $p < .001$, $\chi^2/df = 2.100$, RMSEA = 0.061, SRMR = 0.033, CFI = .999, and TLI = .999. Factor loadings and residuals were given on the path diagram in Figure 2.

Figure 2. *Second-order CFA path diagram.*



In Figure 2, the abbreviation DALS was used for digital assessment literacy self-efficacy, Factor 1 for using appropriate tool and assessment type, Factor 2 for feedback and developing appropriate tools for the objectives and levels, and Factor 3 for preparing exams and evaluating the usability of the scores. According to Figure 2, factor loadings ranged between .785 and .957 and all items made a significant contribution to the model. When the second-order CFA results were evaluated together, it was seen that the fit of the hypothetically tested model was high, and the construct validity of the scale was at a sufficient level.

3.4. Reliability, Convergent and Divergent Validity

Campbell and Fiske (1959) suggest examining convergent and divergent validity for construct validity. While convergent validity is the degree to which latent variables are adequately measured with the relevant items, divergent validity indicates the degree to which the measurements of different latent variables are not related. To provide further evidence for the construct validity of the scale, average variance extracted (AVE), composite reliability (CR), maximum shared variance (MSV), and average shared variance (ASV) values were calculated for each factor (see Table 5).

Table 5. *Convergent and divergent validity and reliability statistics.*

Factors	AVE *(above .50)*	CR *(above .60)*	MSV	ASV	α	ω	Convergent validity *(CR > AVE AVE > .50)*	Divergent validity *(MSV < AVE ASV < AVE)*
Factor 1	.71	.96	.55	.54	.96	.96	YES	YES
Factor 2	.78	.96	.65	.59	.95	.95	YES	YES
Factor 3	.77	.96	.65	.60	.96	.97	YES	YES

Note. Cronbach's α and McDonald's ω internal consistency values were .98. Composite reliability for all items was .99.

For convergent validity, AVE value should be greater than .50 and smaller than CR value, and CR value should be greater than .60, while for divergent validity, MSV and ASV values should be smaller than AVE value (Hair *et al*., 2009; Fornell & Larcker, 1981). According to Table 5, it was seen that convergent and divergent validity criteria were met. Kline (2016) suggests that a value of .90 and above for a reliability coefficient indicates a perfect level of reliability. When Table 5 was examined, it was seen that the reliability values for all data sets were at high levels and also provided evidence for construct validity.

3.5. Known-Groups Validity

According to Cronbach and Meehl (1955), examining group differences in terms of the measured construct is one of the methods that can be used to determine construct validity. Whether there is a difference between the scores of individuals in terms of receiving training on digital assessment was analyzed with Mann Whitney U test. In the preliminary analyses, it was observed that the skewness and kurtosis values were out of the range of -1 to +1, the histogram graphs showed deviation from normal distribution, and the normality test were found to be significant ($p < .001$). In the outlier analysis, 17 participants were found to have extreme values. After these participants removed, the analysis was performed with 279 participants. The results showed that individuals who received training on digital assessment have higher self-efficacy levels than individuals who did not receive any training ($z = -5.536$, $p = .000$). The results were similar for the factors using appropriate tool and assessment type ($z = -6.139$, $p = .000$), feedback and developing appropriate tools for the objectives and level ($z = -4.791$, $p = .000$) and preparing exams and evaluating the usability of the scores ($z = -4.762$, $p = .000$). These findings revealed that the construct validity of the developed scale was found to be at a sufficient level in terms of group differences.

3.6. Findings for Practice

In order to examine how the DALSS scale yielded results in practice, the total scale scores were compared according to the participants' gender, professional experience, teaching field and teaching level. For all variables, although skewness and kurtosis values were in the (-1,+1) range, histogram graphs were negatively skewed and showed deviation from standard normal distribution. Moreover, normality tests were found to be significant ($p < .001$). Hence, nonparametric analyses were conducted. Differentiation of DALSS scores in terms of gender was analyzed by using the Mann-Whitney U test and the Kruskal-Wallis H test for other variables. The results showed that teachers' digital assessment literacy self-efficacy scores did not differ significantly according to their gender ($z = -1.332$, $p = .183$), professional experience ($\chi^2[4] = 6.256$, $p = .181$), and teaching level ($\chi^2[2] = 1.218$, $p = .544$). However, they differed significantly according to their teaching field ($\chi^2[6] = 18.896$, $p = .004$). Primary, foreign language, and other branches (Vocational, art, philosophy, informatics, etc.) teachers' scores were higher than Turkish and social science teachers' scores. Similarly, foreign language and other branches teachers' scores were higher than science, mathematics and educational science teachers' scores.

4. DISCUSSION and CONCLUSION

In this study, it was aimed to develop a psychometrically adequate measurement tool to determine teachers' digital assessment literacy self-efficacy. For this purpose, the steps of the scale development process were followed and various evidence regarding validity and reliability of the scale was obtained. For the theoretical background of the scale, a literature review was conducted, expert opinions were consulted, and Lawshe (1975) content validity index was calculated. After the content validity of the scale was ensured, a draft scale form that could be scored on a 5-point Likert scale was created. An EFA was performed on the draft form applied to a sufficiently large sample after obtaining the necessary ethical approval and permissions. The results revealed that DALSS was found to have a 3-factor structure consisting of 24 items. The hypothetical structure of the scale was tested by CFA, and it was seen that the theoretical structure adequately fit the data. Second-order CFA also showed that all items and latent variables had a good fit to the data.

The dimensions of the scale were named as “using appropriate tool and assessment type”, “feedback and developing appropriate tools for the objectives and levels”, and “preparing exams and evaluating the usability of the scores” based on the related literature. Due to the complexity of the assessment literacy and its use in digital environments, the 3-factor model supported its multidimensional structure. Parallel to this result, Eyal (2012) stresses the importance of having the ability of effectively using assessment data and digital tools, planning of assessment process, managing assessment scores in a digital environment and interpreting the results for digital assessment literacy. Moreover, assessment literacy includes understanding basic assessment concepts and procedures (Popham, 2011), designing assessment instruments, collecting relevant data with appropriate tools, analyzing and interpreting collected data, and using assessment results in meaningful ways (Yamtim & Wong-Wanich, 2014). Hence, the dimensions of DALSS and their items represented different aspects of assessment literacy in digital environments.

Analyses based on group differences were conducted to provide evidence for the construct validity of the scale. Accordingly, it was found that the scale items were able to distinguish between individuals with high and low DALSS scores. In addition, it was also able to reveal the difference between individuals who received and did not receive training on digital assessment. Reliability analyses showed that the scale's Cronbach's α and McDonald's ω values were at a perfect level for different data sets. The scores that can be obtained from the DALSS range from 24 to 120. The interval of scores for the factor “using appropriate tool and assessment type” ranges from 10 to 50, for “feedback and developing appropriate tools for the objectives and levels” from 6 to 30, and for “preparing exams and evaluating the usability of the scores” from 8 to 40. High scores on the scale indicate that teachers' digital assessment literacy self-efficacy is high. Since DALSS was developed and validated in Turkish language (Appendix A), a translation of the items to the English language were given in the Table 3.

According to the findings of the scale for practice, it was found that DALSS scores did not differentiate according to gender. In the literature, however, some studies comparing digital literacy skills (Cabezas-Gonzalez *et al.*, 2017, Lucas *et al.*, 2021) have revealed that male teachers have higher mean proficiency scores than female teachers. Although digital literacy is found higher in favor of males in the literature, assessment skills might have a different effect on digital literacy. In other words, assessment literacy skills combined with digital literacy need to be evaluated apart from the profession in proficiency in digital environments in general. In addition, teachers' DALSS scores do not differ significantly according to professional experience and teaching level. Mertler (2004) also tried to compare pre-service and in-service teacher assessment literacy and similar to the results of this study and he found that there was no significant relationship between teaching experience and teachers' assessment literacy levels. The fact that why the DALSS scores did not differentiate according to professional experience and teaching level might be a significant area of investigation. However, DALSS scores

differentiate significantly in terms of teaching field. Accordingly, the teachers in the field of foreign languages and other branches such as informatics, vocational sciences and arts have higher DALSS scores than teachers in the field of Turkish language, social sciences, mathematics. This result might stem from the fact that teachers in the field of informatics and vocational sciences become more familiar with digital learning environments. Nevertheless, it seems to be important to examine this difference between teaching fields in a broader sense.

4.1. Conclusions, Limitations, Implications and Suggestions

DALSS is the first attempt to develop a scale to assess teachers' self-efficacy for assessment in digital environments. Our validation study contributes to addressing the need for a valid measurement tool to evaluate teachers' digital assessment self-efficacy. Therefore, DALLS will help to measure teachers' digital assessment competencies, provide data to relevant institutions and organizations with the results obtained, and make new decisions regarding education. In addition, it can be suggested to use this scale on different samples of teachers to address the digital assessment competencies of teachers from a broader perspective. There is no drawback for researchers to use the scale in various applications to be made for this purpose.

We acknowledge that more international data are needed to further validate the research results for other countries since DALSS were developed with teachers from Turkey. In addition, the fact that the study did not collect data based on observation and interviews constitutes another limitation of the study.

The study was based on data obtained from teachers working in various branches and grade levels in Turkey. In future studies, studies on a larger sample of teachers from other countries can be conducted. In addition, more concrete results can be obtained by integrating the data obtained as a result of the application of the scale with the data based on observation and interviews. Also, it should investigate the relationships between teachers' self-efficacy for assessment in digital environments and teachers' digital competencies, beliefs, and attitudes toward technology.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Necmettin Erbakan University Social Sciences and Humanities Scientific Research Ethics Committee, 12603 (Approval date: 20/01/2023).

Contribution of Authors

Ruhan Karadağ Yılmaz: Investigation, Resources, Methodology, Supervision, Validation and Writing-original draft. **İlhan Koyuncu**: Investigation, Resources, Methodology, Validation, Visualization, Software, Formal Analysis, and Writing-original draft.

Orcid

Ruhan Karadağ Yılmaz https://orcid.org/0000-0003-3254-8890
İlhan Koyuncu https://orcid.org/0000-0002-0009-5279

REFERENCES

American Federation of Teachers [AFT], National Council on Measurement in Education [NCME], & National Education Association [NEA] (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice, 9*(4), 30-32.

Al-Bahlani, S.M. (2019). *Assessment literacy: A study of EFL teachers' assessment knowledge, perspectives, and classroom behaviors* [Doctoral dissertation, The University of Arizona]. Available from ProQuest Dissertations & Theses Global (2229389562). https://www.proq

uest.com/dissertations-theses/assessment-literacy-study-efl-teachers-knowledge/docview/2229389562/se-2

Al-Bahlani, S.M., & Ecke, P. (2023). Assessment competence and practices including digital assessment literacy of postsecondary English language teachers in Oman. *Cogent Education, 10*(2), 2239535. https://doi.org/10.1080/2331186X.2023.2239535

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

Boateng, G.O., Neilands, T.B., Frongillo. E.A., Melgar-Quiñonez. H.R., & Young, S.L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Front Public Health. 11*(6), 1-18. https://doi.org/10.3389/fpubh.2018.00149

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154*, 136-136. https://doi.org/10.1177/0049124192021002005

Cabezas-Gonzalez, M., Casillas-Martín, S., Sanches-Ferreira, M., & Teixeira Diogo, F.L. (2017). Do gender and age affect the level of digital competence? A study with university students. *Fonseca Journal of Communication, 15*, 109-125. https://doi.org/10.14201/fjc201715109125

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Cattell, R.B. (1978). The *scientific use of factor analysis in behavioral and life sciences*. Plenum Press. https://doi.org/10.1007/978-1-4684-2262-7

Chan, C.K.Y., & Luo, J. (2020). A four-dimensional conceptual framework for student assessment literacy in holistic competency development. *Assessment & Evaluation in Higher Education, 46*(3), 451-466. https://doi.org/10.1080/02602938.2020.1777388

Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis*. Lawrence Erlbaum.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin 52*(4), 281. https://doi.org/10.1037/h0040957

DeLuca, C., & Klinger, D.A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17*(4), 419-438. https://doi.org/10.1080/0969594X.2010.516643

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016a). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. *Educational Assessment, 21*(4), 248-266. https://doi.org/10.1080/10627197.2016.1236677

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016b). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation, and Accountability, 28*, 251–272. https://doi.org/10.1007/s11092-015-9233-6

Edwards, F. (2016). A rubric to track the development of secondary pre-service and novice teachers' summative assessment literacy. *Assessment in Education: Principles, Policy & Practice, 24*(2), 1-23. https://doi.org/10.1080/0969594X.2016.1245651

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme [Measurement and scale development in Psychology]*. Pegem Akademi.

Eyal, L. (2012). Digital assessment literacy - the core role of the teacher in a digital environment. *Educational Technology & Society, 15*(2), 37-49. https://www.jstor.org/stable/jeductechsoci.15.2.37

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2011). *How to design and evaluate research in education*. McGraw-Hill.

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly,* 9, 113–132. https://doi.org/10.1080/15434303.2011.642041

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2009). *Multivariate data analysis*. Prentice-Hall.

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, 6*, 1-55. https://doi.org/10.1080/10705519909540118

Kline, R.B. (2016). *Principle and practice of structural equation modeling*. The Guilford Press.

Koyuncu, I., & Kılıç, A. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Education and Science*, *44*(198), 361-388.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lucas, M., Bem-Haja, P., Siddiq, F., Moreira, A., & Redecker, C. (2021). The relation between in-service teachers' digital competence and personal and contextual factors: What matters most? *Computers & Education, 160*, 104052. https://doi.org/10.1016/j.compedu.2020.104052.

Mertler, C.A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 33*(1), 49-64. https://www.jstor.org/stable/41064623

Mertler, C.A., & Campbell, C.S. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory* (Paper presentation). The annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Ministry of National Education [MoNE] (2017). *General qualifications for teaching profession*. General Directorate of Teacher Training and Development. https://oygm.meb.gov.tr/dosyalar/StPrg/Ogretmenlik_Meslegi_Genel_Yeterlikleri.pdf

Mindrila, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, *1*(1), 60-66.

Morgado, F.F., Meireles, J.F., Neves, C.M., Amaral, A.C., & Ferreira, M.E. (2018). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica, 30*(3), 1-20. https://doi.org/10.1186/s41155-016-0057-1

Nimehchisalem, V., & Bhatti, N. (2019). A review of literature on language assessment literacy in last two decades (1999-2018). *International Journal of Innovation, Creativity and Change, 8*(11), 44-59.

Nyagi, K., & Rajendran, M. (2020). Pre-service teachers' approaches to classroom assessment. *Humanities & Social Sciences Reviews, 8*(1), 666-673. https://doi.org/10.18510/hssr.2020.8180

Plake, B.S., & Impara, J.C. (1992). *Teacher competencies questionnaire description*. University of Nebraska.

Popham, W.J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*, 265–273. https://doi.org/10.1080/08878730.2011.605048

R Core Team (2023). R: A Language and environment for statistical computing. (Version 4.3) [Computer software]. https://cran.r-project.org. (R packages retrieved from CRAN snapshot 2024-01-09).

Rezai, A., Alibakhshi, G., Farokhipour, S., & Miri, M. (2021). A phenomenographic study on language assessment literacy: hearing from Iranian university teachers. *Language Testing in Asia, 11*(26), 1-25. https://doi.org/10.1186/s40468-021-00142-5

Rogier, D. (2014). Assessment literacy: Building a base for better teaching and learning. *English Teaching Forum, 52*(3), 2-13. https://americanenglish.state.gov/files/ae/resource_files/etf_52_3_02-13.pdf

Rosseel, Y. (2019). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Shams, J.A., Iqbal, M.Z., & Iqbal, M.Z. (2018). Investigation of classroom assessment literacy of university teachers of Punjab. *Pakistan Journal of Distance and Online Learning, IV*(II), 103-118. https://files.eric.ed.gov/fulltext/EJ1267033.pdf

Stiggins, R. (1991). Assessment literacy. *The Phi Delta Kappan, 72*, 534–539. https://learnline.cdu.edu.au/commonunits/documents/Scaffolding%20and%20formative%20assessment.pdf

Stiggins, R.J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23-17. https://doi.org/10.1111/j.1745-3992.1999.tb00004.x

Tabachnick, B.G., & Fidell, L.S. (2012). *Using multivariate statistics*. Allyn & Bacon.

The jamovi project (2024). Jamovi (Version 2.5) [Computer Software]. https://www.jamovi.org

Ullman, J.B. (2012). Structural equation modeling. In B.G. Tabachnik & L.S. Fidell (Eds.), *Using multivariate statistics* (6th ed.). Pearson.

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*(3), 749-770. https://eric.ed.gov/?id=EJ780818

Walters, F.S. (2010). Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly, 7*(4), 317-342. https://doi.org/10.1080/15434303.2010.516042

Worthington, R.L., & Whittaker, T.A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838. https://doi.org/10.1177/0011000006288127

Yamtim, V., & Wong-Wanich, S. (2014). A study of classroom assessment literacy of primary school teachers. *Procedia-Social and Behavioral Sciences, 116*, 2998-3004. https://doi.org/10.1016/j.sbspro.2014.01.696

Yong, A.G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology,* 9(2), 79-94. https://doi.org/10.20982/tqmp.09.2.p079

Yuan, K.H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis, 100*(9), 1900-1918.

Zhang, Z., & Burry-Stock, J.A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323-342. https://doi.org/10.1207/S15324818AME1604_4

APPENDICES

Appendix A. Digital Assessment Literacy Self-Efficacy Scale (Turkish Version)

Items	Completely	Disagree (2)	Somewhat	Agree (4)	Completely
1. Değerlendirme sürecinde bilgi ve iletişim teknolojilerini (bilgisayar, tablet, projeksiyon, tarayıcı vb.) kullanabilirim.	1	2	3	4	5
2. Öğrencileri değerlendirirken dijital değerlendirme araçlarını (hesaplama ve çizim araçları, e-ortamlar, paket programlar, özel değerlendirme yazılımları vb.) kullanabilirim.	1	2	3	4	5
3. Çevrimiçi ortamlarda tanılayıcı (hazır bulunuşluk, ön bilgiler vb.) değerlendirme uygulamaları yapabilirim.	1	2	3	4	5
4. Çevrimiçi ortamlarda biçimlendirici (izleme, konu tarama, quizler vb.) değerlendirme uygulamaları yapabilirim.	1	2	3	4	5
5. Çevrimiçi ortamlarda özetleyici (düzey belirleme, not verme vb.) değerlendirme uygulamaları yapabilirim.	1	2	3	4	5
6. Çevrimiçi ortamlarda mutlak değerlendirme uygulamaları yapabilirim.	1	2	3	4	5
7. Çevrimiçi ortamlarda bağıl değerlendirme uygulamaları yapabilirim.	1	2	3	4	5
8. Dijital değerlendirme araçlarıyla öğrencileri değerlendirme süreçlerine dahil edebilirim.	1	2	3	4	5
10. Değerlendirme sürecinde çevrimiçi öğrenme sistemlerindeki (mergen, moodle, eba vb.) öğrenen analitiklerine yer verebilirim.	1	2	3	4	5
11. Değerlendirme süreciyle ilgili yeni dijital teknolojileri takip ederim.	1	2	3	4	5
14. Dersin bilişsel hedeflerine uygun değerlendirme araçlarını dijital ortamlarda hazırlayabilirim.	1	2	3	4	5
15. Duyuşsal becerileri ölçmeye uygun dijital ölçme araçları tasarlayabilirim.	1	2	3	4	5
16. Dijital ortamlarda devinimsel (psikomotor) amaçlara yönelik ölçme araçları geliştirebilirim.	1	2	3	4	5
17. Farklı öğrenme düzeyleri için dijital değerlendirme araçları hazırlayabilirim.	1	2	3	4	5
18. Öğrencilerin gelişimsel özelliklerine uygun değerlendirme araçlarını dijital ortamlarda hazırlayabilirim.	1	2	3	4	5
19. Dijital ortamda elde ettiğim ölçme ve değerlendirme sonuçlarına göre tüm paydaşlara geribildirim verebilirim.	1	2	3	4	5
22. Dijital ortamlarda ölçme ve değerlendirme etkinlikleri (e-sınav, kısa sınav vb.) düzenleyebilirim.	1	2	3	4	5
23. Dijital ortamda elde ettiğim ölçme sonuçlarının geçerliğini (kapsam, yapı, ölçüt) belirleyebilirim.	1	2	3	4	5
24. Dijital ortamda elde ettiğim ölçme sonuçlarının güvenirliğini (iç tutarlılık, test-tekrar test vb.) belirleyebilirim.	1	2	3	4	5
25. Çevrimiçi ortamda farklı değerlendirme araçları için uygun puanlama yöntemlerini (seçmeli, dereceli puanlama, doğru/yanlış vb.) kullanabilirim.	1	2	3	4	5
26. Dijital ortamda farklı değerlendirme araçlarından elde ettiğim bulguları birlikte değerlendirebilirim.	1	2	3	4	5
27. Çevrimiçi değerlendirme araçlarının birbirine göre üstünlük ve sınırlılıklarını karşılaştırabilirim.	1	2	3	4	5
28. Dijital öğrenme ortamlarının yapısına (çevrimiçi, çevrimdışı vb.) uygun değerlendirme araçları geliştirebilirim.	1	2	3	4	5
29. Dijital değerlendirme araçlarının kullanışlılığını (emek, zaman, maliyet vb.) değerlendirebilirim.	1	2	3	4	5

Appendix B. Digital Assessment Literacy Self-Efficacy Scale Draft Items (Turkish version)

1. Değerlendirme sürecinde bilgi ve iletişim teknolojilerini (bilgisayar, tablet, projeksiyon, tarayıcı vb.) kullanabilirim.
2. Öğrencileri değerlendirirken dijital değerlendirme araçlarını (hesaplama ve çizim araçları, e-ortamlar, paket programlar, özel değerlendirme yazılımları vb.) kullanabilirim.
3. Çevrimiçi ortamlarda tanılayıcı (hazır bulunuşluk, ön bilgiler vb.) değerlendirme uygulamaları yapabilirim.
4. Çevrimiçi ortamlarda biçimlendirici (izleme, konu tarama, quizler vb.) değerlendirme uygulamaları yapabilirim.
5. Çevrimiçi ortamlarda özetleyici (düzey belirleme, not verme vb.) değerlendirme uygulamaları yapabilirim.
6. Çevrimiçi ortamlarda mutlak değerlendirme uygulamaları yapabilirim.
7. Çevrimiçi ortamlarda bağıl değerlendirme uygulamaları yapabilirim.
8. Dijital değerlendirme araçlarıyla öğrencileri değerlendirme süreçlerine dahil edebilirim.
9. Çevrimiçi ortamlarda öğrencileri değerlendirmek amacıyla veri toplayabilirim.
10. Değerlendirme sürecinde çevrimiçi öğrenme sistemlerindeki (mergen, moodle, eba vb.) öğrenen analitiklerine yer verebilirim.
11. Değerlendirme süreciyle ilgili yeni dijital teknolojileri takip ederim.
12. Çevrimiçi ölçme araçları ile elde edilen öğrenci puanları üzerinde betimleyici istatistikleri (ortalama, standart sapma, mod, medyan vb.) hesaplayabilirim.
13. Çevrimiçi ölçme araçları ile elde edilen öğrenci puanlarını tablo ve grafiklerle gösterebilirim.
14. Dersin bilişsel hedeflerine uygun değerlendirme araçlarını dijital ortamlarda hazırlayabilirim.
15. Duyuşsal becerileri ölçmeye uygun dijital ölçme araçları tasarlayabilirim.
16. Dijital ortamlarda devinimsel (psikomotor) amaçlara yönelik ölçme araçları geliştirebilirim.
17. Farklı öğrenme düzeyleri için dijital değerlendirme araçları hazırlayabilirim.
18. Öğrencilerin gelişimsel özelliklerine uygun değerlendirme araçlarını dijital ortamlarda hazırlayabilirim.
19. Dijital ortamda elde ettiğim ölçme ve değerlendirme sonuçlarına göre tüm paydaşlara geribildirim verebilirim.
20. Dijital ortamda tamamlayıcı ölçme ve değerlendirme araçlarını (portfolyo, kavram haritaları, gözlem formları, öz ve akran değerlendirme vs.) uygulayabilirim.
21. Dijital ortamlarda öğrencilerin bireysel çalışmalarını (ödev, proje vb.) değerlendirebilirim.
22. Dijital ortamlarda ölçme ve değerlendirme etkinlikleri (e-sınav, kısa sınav vb.) düzenleyebilirim.
23. Dijital ortamda elde ettiğim ölçme sonuçlarının geçerliğini (kapsam, yapı, ölçüt) belirleyebilirim.
24. Dijital ortamda elde ettiğim ölçme sonuçlarının güvenirliğini (iç tutarlılık, test-tekrar test vb.) belirleyebilirim.
25. Çevrimiçi ortamda farklı değerlendirme araçları için uygun puanlama yöntemlerini (seçmeli, dereceli puanlama, doğru/yanlış vb.) kullanabilirim.
26. Dijital ortamda farklı değerlendirme araçlarından elde ettiğim bulguları birlikte değerlendirebilirim.
27. Çevrimiçi değerlendirme araçlarının birbirine göre üstünlük ve sınırlılıklarını karşılaştırabilirim.
28. Dijital öğrenme ortamlarının yapısına (çevrimiçi, çevrimdışı vb.) uygun değerlendirme araçları geliştirebilirim.
29. Dijital değerlendirme araçlarının kullanışlılığını (emek, zaman, maliyet vb.) değerlendirebilirim.
30. Dijital ortamlardaki değerlendirme sürecini dersin hedeflerine uygun olacak şekilde düzenleyebilirim.
31. Çevrimiçi ortamda etik ilkeleri göz önünde bulundururum.