

Turkish Validity and Reliability of Peer Evaluation Scale in Team-Based Learning

ABSTRACT

Objective: This study was conducted to adapt the “Peer Evaluation Scale in Team-Based Learning” into Turkish for use with students enrolled in health programs. Participants included students from four distinct health programs.

Methods: This research was conducted as a methodological study. Following the pilot study (n = 117), the scale was administered to a larger sample for the main psychometric analyses, yielding data from 590 participants. Data were collected using the “Personal Information Form,” the draft “Peer Evaluation Scale for Team-Based Learning,” and “Visual Analog Scale for Peer Evaluation”. The data were analyzed using the content validity index, confirmatory factor analysis, and reliability analysis.

Results: The fit index values of the scale were $\chi^2/df=4.617$, GFI=0.936, AGFI=0.900, CFI=0.950, RMSEA=0.078, RMR=0.019, SRMR=0.0382, NFI=0.937, TLI=0.933, IFI=0.950, PGFI=0.600, PNFI=0.710. Cronbach’s Alpha values were 0.832 for the “Responsibility” sub-dimension, 0.808 for the “Initiative” sub-dimension, and 0.757 for the “Cooperation” sub-dimension. The Cronbach’s Alpha for the total scale was 0.921. The Turkish version of the 12-item, 3-subdimensional scale was validated without any alterations to the original scale structure.

Conclusion: The Turkish version of the Peer Review Scale for Team-based Learning demonstrates validity and reliability for use in Turkish populations. These findings support the use of the Turkish version of the Peer Review Scale for Team-Based Learning as a valid and reliable tool for assessing peer evaluation in team-based learning settings in both research and educational practice.

Keywords: Learning, peer evaluation, reliability, teamwork, Turkish adaptation, validity

Abdulsamed KAYA¹

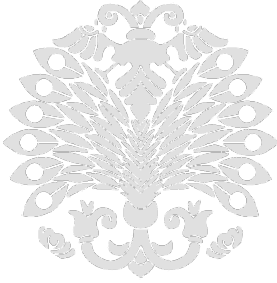


¹Mus Alparslan University, Department of Medical Services and Techniques, Vocational School of Health Services, Mus, Türkiye

Güzel Nur YILDIZ¹



¹Mus Alparslan University, Department of Medical Services and Techniques, Vocational School of Health Services, Mus, Türkiye



Received Date 10.07.2025
Accepted Date 18.12.2025
Publication Date 30.03.2026

Corresponding author: Abdulsamed KAYA

E-mail: kaya.abdulsamed@gmail.com

Cite this article: Kaya, A., & Yıldız, G.N., (2025). Turkish Validity and Reliability of Peer Evaluation Scale in Team-Based Learning. *Journal of Midwifery and Health Sciences*, 9(1), 16-26.



Content of this journal is licensed under a Creative Commons Attribution-Noncommercial 4.0 International License.

Introduction

Numerous health-related organizations emphasize the importance of interprofessional collaboration to address the multifaceted challenges faced by health systems. The ability to collaborate effectively is essential in healthcare organizations due to their inherently complex structures (Whitley et al., 2015). To align with this trend, Team-Based Learning (TBL) has been proposed in health education as a method to facilitate the development of collaborative skills among students (Rezaee & Moadeb, 2016). In this context, team-based practices are implemented to train future nursing professionals in problem-solving and collaborative competencies (Zwarenstein et al., 2009).

TBL is recognized as an impactful approach for fostering healthcare professionals' learning processes, interpersonal interactions, decision-making skills, and team coordination abilities (Alberti et al., 2021). In addition, research shows that TBL contributes to improved academic outcomes among students and helps overcome typical difficulties experienced within group dynamics (Rotgans et al., 2018). TBL is an active learning approach designed to facilitate teamwork skills while achieving course objectives. Many faculty members have adopted TBL due to its distinctive pedagogical approach (Sisk, 2011). In TBL, teams are formed comprising small groups of five to seven students. Through collective decision-making, such teams foster problem-solving competencies and sustain their development within a cohesive structure (Parmelee & Michaelsen, 2010).

The TBL methodology is characterized by the following features: (a) Most classroom sessions are dedicated to students working in stable, deliberately structured teams; (b) students achieve conceptual understanding by preparing individually before class, followed by an in-class readiness check that includes both solo and group evaluations; (c) instructional time is predominantly spent on team-based tasks, encouraging collaborative engagement with practical aspects of the course content; (d) peer review and feedback mechanisms ensure students remain accountable to one another. Moreover, TBL facilitates the transformation of groups from a collection of individuals to high-performing and self-managed learning teams (Michaelsen & Sweet, 2011). TBL participants gain valuable experience that helps them build leadership and interpersonal abilities. It also contributes to their growth in problem-solving, motivation, and personal responsibility (Alberti et al., 2021).

Nursing programs incorporate TBL as a strategy to equip students with the competencies required to navigate multifaceted clinical situations, emphasizing collaborative interactions and effective dialogue with both patients and healthcare teams (Kweon & Park, 2023). Among the criticisms of TBL are the inconsistent allocation of work

within groups and the possibility that reliance on high-performing students may inhibit active learning for all participants (Michaelsen et al., 2024).

Among the methods employed to involve students in TBL assessment, peer assessment is the most common. This technique reveals important aspects of team dynamics and learning progression that often remain invisible to educators. Moreover, it contributes to the development of students' cognitive and affective abilities, including their problem-solving skills, objective critique of work, awareness of peers' contributions, and sense of responsibility (Levine & Hudes, 2021). Through receiving evaluations from their peers, health professionals can pinpoint weaknesses, refine their expertise, and ultimately improve patient care quality (Burgess et al., 2020). Consequently, peer review within health education serves as a valuable means for healthcare staff to develop proficient feedback skills. Despite the broad recognition of peer assessment's significance in learner-centered approaches, numerous challenges still need to be addressed (Kweon & Park, 2023).

TBL represents a rigorously organized, evidence-supported instructional approach. Nonetheless, considerable variation exists in how peer assessments are applied (Considine et al., 2021). When assessments are inaccurate, they may negatively affect the effectiveness of learning (Goedde & Sibley, 2014). Furthermore, these scales were constructed according to the requirements of researchers and trainers. Consequently, they may not fully address the needs of the learners (Kweon & Park, 2023). Generally, it has been observed that students express satisfaction with TBL, and student engagement is elevated in TBL classrooms. Additionally, evidence suggests that students in TBL classes demonstrate superior performance on examinations (Sisk, 2011).

In Türkiye, the use of team-based learning has been increasingly adopted in health-related education programs, emphasizing the need for systematic and contextually appropriate assessment tools. Peer evaluation is a core component of TBL; however, students' perceptions of responsibility, initiative, and cooperation may be shaped by cultural and educational norms. Despite this growing implementation, there is currently no standardized, validated Turkish-language instrument specifically designed to assess peer evaluation within TBL settings. Existing measures are either not specific to the TBL framework. Therefore, adapting a well-established TBL-specific peer evaluation scale into Turkish is essential to ensure culturally relevant, reliable, and valid assessment practices in national educational contexts.

In the national literature, there is an absence of a measurement instrument for TBL and peer evaluation.

Therefore, this study aimed to adapt the “Peer Evaluation Scale for TBL,” which was developed by Kweon and Park in 2023 with 12 items (three dimensions), into Turkish.

Hypotheses:

H1: The three-factor structure of the Peer Evaluation Scale for Team-Based Learning will be confirmed in the Turkish sample, indicating adequate construct validity.

H2: The Peer Evaluation Scale for Team-Based Learning and its subscales will demonstrate acceptable internal consistency reliability, with Cronbach’s Alpha coefficients ≥ 0.70 .

H3: Scores obtained from the Peer Evaluation Scale for Team-Based Learning will show significant positive correlations with scores from the Visual Analog Scale for Peer Evaluation, providing evidence for criterion-related validity.

Methods

Type of the Study

This investigation constitutes a methodologically designed scale adaptation study. In line with the International Test Commission (ITC) guidelines, the adaptation process aimed to ensure linguistic and cultural equivalence, while construct equivalence was evaluated through confirmatory factor analysis.

Place and Time of the Study

The research was conducted with students enrolled in “Elderly Care,” “Disabled Care and Rehabilitation,” “Nursing,” and “First and Emergency Aid” programs at Muş Alparslan University. To implement the research, students were initially briefed on TBL, and teams comprising five to seven members were formed based on class size. Between April and May 2024, the teams were assigned TBL-related tasks and activities. These activities were conducted both within and outside the classroom setting. Upon completion of the two-month TBL process, data collection was carried out in June 2024. Data were gathered utilizing “Google Forms.” Students evaluated their team members, resulting in each student completing the questionnaire between four and six times.

Participants and Sampling Method

The study population comprised students aged 18 and above, enrolled in the “Elderly Care,” “Disabled Care and Rehabilitation,” “Nursing,” and “First and Emergency Aid” programs. Participants were selected using a convenience sampling method, and participation was based on voluntary consent among students enrolled in the selected classes. Data for the pilot study were collected from the “First and Emergency Aid” program, while data for the main study

were obtained from the “Elderly Care,” “Disabled Care and Rehabilitation,” and “Nursing” programs. The pilot implementation phase yielded 117 data points, and the main study yielded 590 data points. In the team-based learning (TBL) context, each student evaluated all other members of their team individually. Therefore, although 169 students participated in the study, a total of 590 peer evaluation forms (data points) were obtained, as each participant provided multiple peer assessments. Each peer evaluation form was treated as a unit of analysis, consistent with previous peer assessment studies in TBL contexts, as the focus of the analysis was the psychometric properties of the scale rather than individual-level outcomes. Sample adequacy for confirmatory factor analysis was evaluated based on recommendations in the literature. In addition, rules of thumb such as having at least 10 participants per item or per factor have been commonly applied. In the present study, the scale consisted of 12 items and three factors, and data were obtained from 169 students, exceeding these recommended thresholds. Moreover, due to the team-based learning context, a total of 590 peer evaluation forms were collected, further supporting the adequacy of the sample size for Confirmatory Factor Analysis (CFA) (Seçer, 2020).

Inclusion criteria

18 years of age or older, participation in research, inclusion in a group in the TBL process and not to be absent during the research process.

Exclusion criteria

Non-participation or withdrawal from the study, absenteeism or dropping out of school and incomplete or incorrect filling of data.

Data Collection

The data collection instruments employed in this study comprised the “Personal Information Form,” the draft “Peer Evaluation Scale for Team-Based Learning,” and the “Visual Analog Scale for Peer Evaluation.”

Personal Information Form

This instrument comprises inquiries regarding the participants' age, gender, and their perception of the TBL as educational and instructive.

Peer Evaluation Scale for Team-Based Learning

This scale was developed by Kweon and Park in 2023 to facilitate peer evaluation in TBL (Kweon & Park, 2023). The scale comprises 12 items and three sub-dimensions. The instrument comprises three distinct sub-dimensions: responsibility (items 2, 4, 5, 7), initiative (items 1, 8, 9, 12), and cooperation (items 3, 6, 10, 11). It is designed as a 5-

point Likert-type scale, allowing respondents to indicate their level of agreement with each statement, ranging from 1 = strongly disagree to 5 = strongly agree. The Cronbach's Alpha values for the responsibility, initiative, and cooperation sub-dimensions were 0.840, 0.791, and 0.802, respectively (Kweon & Park, 2023).

Visual Analog Scale for Peer Evaluation

Participants were asked to rate each team member using a numerical scale ranging from 0 (lowest possible score) to 100 (highest possible score). Correlation analyses were conducted between the obtained values and the scale and subdimensions.

Adaptation Process

Adaptation of the measurement tool was conducted through a structured three-phase approach.

First stage (translation of the items from English to Turkish, submission of the items to expert opinion, and preparation for pilot application): The initial translation of the scale into Turkish was conducted independently by two English language experts. Following the translation, the Turkish version was submitted for evaluation by a panel of 12 specialists. This panel consisted of 4 nurses with expertise in Internal Medicine and Care, 4 experts in Nursing Principles, 1 nurse specializing in Obstetrics and Gynecology, 1 expert in Pediatric Nursing and Care, and 2 Turkish language professionals. Content validity was assessed using the Davis technique based on the Content Validity Index (CVI) (Davis, 1992). A panel of subject-matter experts was asked to rate each item in terms of relevance using a 4-point Likert scale (1 = not relevant, 4 = highly relevant). Item-level Content Validity Index (I-CVI) was calculated as the proportion of experts who rated an item as either 3 or 4. Items with an I-CVI value of 0.80 or higher were retained, in line with the universal agreement (UA) criterion (Davis, 1992). Since none of the items scored below the 0.80 threshold, no item was excluded from the scale. Adjustments were carried out according to expert recommendations, followed by a back-translation of the scale into English. Semantic analysis between the first and final drafts confirmed consistency. The updated version of the scale was transferred to Google Forms in order to conduct a pilot test.

Second stage (pilot application): During the pilot application phase, data were collected exclusively from students enrolled in First and Emergency Aid programs. This initial stage yielded 117 data points from 22 participants. The item correlation values for the 12 items ranged from 0.702 to 0.900. The scale's 12 items yielded a Cronbach's Alpha of 0.965. Since this value was well above the 0.70 threshold, none of the items were removed during the pilot phase, allowing the full set of items to be included in the main study

(DeVellis & Thorpe, 2021; Seer, 2020; Shrestha, 2021).

Third stage (Main application): In the third stage (main application), the items were entered into "Google Forms", and a corresponding link was created for data collection. The link was disseminated to students enrolled in "Elderly Care," "Disabled Care and Rehabilitation," and "Nursing programs". Subsequent to gathering 590 data points from 169 participants, confirmatory factor analysis and reliability evaluations were carried out. The CFA fit indices were assessed in accordance with scholarly references (Bae, 2017; Woo, 2017).

Statistical Analysis

The questionnaire items were entered into Google Forms with data protection measures activated to safeguard against data loss. Once responses were collected, the dataset was exported to an Excel file and subsequently transferred into SPSS for statistical processing. The normality of the data distribution was assessed using skewness and kurtosis values (+1/-1). CFA was carried out using the AMOS software. Item-total correlation coefficients exceeding 0.30 and Cronbach's Alpha values above 0.60 were considered indicators of item adequacy. To assess the adequacy of the dataset and sample size, the Kaiser-Meyer-Olkin (KMO) statistic was examined (values above 0.60 were considered acceptable), along with Bartlett's test of sphericity (significant at $p < .05$) and anti-image correlations, with 0.50 set as the minimum threshold (Pallant, 2020; Seer, 2020). Construct validity was assessed via CFA using AMOS, with fit indices interpreted in line with established literature (Bae, 2017; Erkorkmaz et al., 2013; Kline, 2014). To evaluate the internal consistency of the scale, both Cronbach's Alpha and McDonald's Omega coefficients were calculated. Coefficients above the 0.60 threshold were interpreted as indicators of acceptable reliability. Additionally, parallel-form (criterion-related) validity was evaluated using the Visual Analog Scale (VAS) for Peer Evaluation, and the correlation between this measure and the primary scale was used as supporting evidence of validity. The discriminative validity of the scale was analyzed by comparing the mean scores of participants in the lowest and highest 27% of the distribution through an independent samples t-test. A statistically significant difference between these two groups demonstrated that the scale effectively distinguishes between individuals with varying levels of the measured trait (Pallant, 2020; Seer, 2020).

Ethical Considerations and Permissions

Permission to adapt the original scale into Turkish was formally obtained from its developer. Ethical approval for the study was granted by the Scientific Research and Publication Ethics Committee of Muş Alparslan University on February 6, 2024 (Approval No: 129968), in accordance with

the ethical principles outlined in the Declaration of Helsinki. Furthermore, institutional authorization was received for the implementation of the research (Document No: 135120). Prior to data collection, participants were fully informed about the study's purpose, and their verbal consent was obtained voluntarily.

Results

Results Related to Participants

During the primary implementation phase, data were collected from a total of 169 students. The results indicated that 41.4% of the students were enrolled in the nursing program, 74.6% were female, 88.2% found TBL engaging, 93.5% found TBL educational, and 52.7% were within the 18-20 age range (Table 1).

Results Related to Validity

Content Validity Index (CVI)

Based on evaluations by subject matter experts, the I-CVI ranged from 0.80 to 1.00. As a result, all items were retained in the scale without any removals.

Table 1. <i>Sociodemographic profile of the participants</i>			
Features	Variables	N	%
Program	Nursing	70	41.4
	Disabled care and rehabilitation	40	23.7
	Elderly care	59	34.9
Gender	Woman	126	74.6
	Male	43	25.4
Finding TBL enjoyable	Yes	149	88.2
	No	20	11.8
Finding TBL instructive	Yes	158	93.5
	No	11	6.5
Age	18-20	89	52.7
	21 and above	80	47.3

Structure Validity: The suitability of the data set for factor analysis and the adequacy of the sample size were assessed using the KMO coefficient, Bartlett's sphericity test, and anti-image correlations. The KMO result was 0.949, indicating excellent sampling adequacy, and Bartlett's test produced a significant result ($\chi^2=3624.740$, $p=.001$). Antiimage correlations were observed to range from 0.929 to 0.965. No extreme values were detected in the dataset, and the data were determined to be normally distributed.

These findings confirm that the dataset meets the necessary criteria for proceeding with the intended statistical analyses. Item-level analyses showed relatively high mean scores

Table 2.
Item means, correlation coefficients, and Confirmatory Factor Analysis results

Scale Items	Mean $\pm$ SD	Item total correlation value	Cronbach's Alpha value if the item is deleted	Factor Load Values		
				F1	F2	F3
2	4.21 $\pm$ 0.73	0.739	0.911	.79		
4	4.19 $\pm$ 0.77	0.673	0.914	.71		
5	4.25 $\pm$ 0.73	0.674	0.914	.72		
7	4.13 $\pm$ 0.71	0.710	0.912	.75		
1	4.19 $\pm$ 0.73	0.699	0.913		.74	
8	4.18 $\pm$ 0.71	0.684	0.914		.70	
9	4.18 $\pm$ 0.73	0.733	0.911		.75	
12	4.20 $\pm$ 0.72	0.647	0.915		.66	
3	4.18 $\pm$ 0.69	0.686	0.913			.72
6	4.28 $\pm$ 0.67	0.642	0.915			.65
10	4.21 $\pm$ 0.69	0.659	0.915			.69
11	4.13 $\pm$ 0.70	0.504	0.921			.51

F1: Responsibility F2: Initiative F3: Cooperation SD: Standard deviation

across all items (range = 4.1–4.2), indicating a potential ceiling effect. In scale adaptation studies, CFA is recommended to preserve the original structure of the scale. Therefore, CFA was performed to validate the construct structure of the 12-item scale comprising three subdimensions. All factor loadings of the items were above 0.30, ranging between 0.51 and 0.79, which confirms the validity of the proposed 12-item, three-factor model (Table 2).

Confirmatory Factor Analysis (CFA): The CFA results demonstrate that the scale with 12 items and three subdimensions is sufficient to ensure construct validity (Table 3). The PATH diagram generated during CFA is presented in Figure 1.

Results Related to Reliability

Table 4 reveals statistically meaningful correlations among the various sub-dimensions of the scale, underscoring the multidimensional coherence of the instrument. The moderate correlations obtained indicated the absence of a multicollinearity problem among the subdimensions. To assess the reliability of the scale, both Cronbach's Alpha and McDonald's Omega values were utilized.

Table 3.
Confirmatory factor analysis results

Index	Normal Value	Acceptable Value	Measurement	Conclusion
χ^2 'p' Value	$p > .05$	-	0.001	Perfect fit
χ^2/df (CMIN/DF)	<2	<5	4.617	Acceptable fit
GFI	>0.95	>0.90	0.936	Acceptable fit
AGFI	>0.95	>0.85	0.900	Acceptable fit
CFI	>0.95	>0.90	0.950	Perfect fit
RMSEA	<0.05	<0.08	0.078	Acceptable fit
RMR	<0.05	<0.08	0.019	Perfect fit
SRMR	<0.05	<0.08	0.0382	Perfect fit
NFI	>0.95	>0.80	0.937	Acceptable fit
TLI	$0.95 < TLI < 1$	$0.90 < TLI < 0.94$	0.933	Acceptable fit
IFI	>0.90	-	0.950	Perfect fit
PGFI	>0.89	>0.50	0.600	Acceptable fit
PNFI	>0.89	>0.50	0.710	Acceptable fit

χ^2 'p' Value: Chi-Square Test p-value; χ^2 (df): Chi-square (Degrees of Freedom); GFI: Goodness of Fit Index; AGFI: Adjusted Goodness of Fit Index; CFI: Comparative Fit Index; RMSEA: Root Mean Square Error of Approximation; RMR: Root Mean Square Residual; SRMR: Standardized Root Mean Square Residual; NFI: Normed Fit Index; TLI: Tucker-Levis Index; IFI: Incremental Fit Index; PGFI: Parsimony Goodness of Fit Index; PNFI: Parsimony Normed Fit Index

The internal consistency of the scale was assessed using both Cronbach's Alpha and McDonald's Omega coefficients. For F1, Cronbach's Alpha was calculated as 0.832, while McDonald's Omega was 0.831. In F2, both reliability coefficients were 0.808. For F3, the Cronbach's Alpha was 0.757 and McDonald's Omega was 0.758. Regarding the entire scale, both coefficients indicated high reliability, with values of 0.921 (Table 4). In addition, significant positive correlations were observed between the total scale and its subdimensions and the "VAS for Peer Evaluation," further supporting the scale's reliability (Table 4).

Determination of the Distinctiveness of the Scale

To evaluate the discriminative capacity of the scale and its subdimensions, an independent samples t-test was

performed by comparing the mean scores of participants in the lower 27% and upper 27% groups. The analysis revealed statistically significant differences between these groups across both the total scale and each subdimension ($p < .05$). These results provide evidence that the scale has adequate discriminatory power (Table 4).

Discussion

This study was conducted to adapt the "Peer Evaluation Scale for TBL," which comprises 12 items and three subdimensions, into Turkish. A comprehensive literature review revealed no measurement tools related to TBL in Türkiye. The adaptation of the Peer Review Scale for Team-Based Learning into Turkish makes an important contribution to the national literature by providing a

Table 4.
Correlation values and mean scores and reliability results

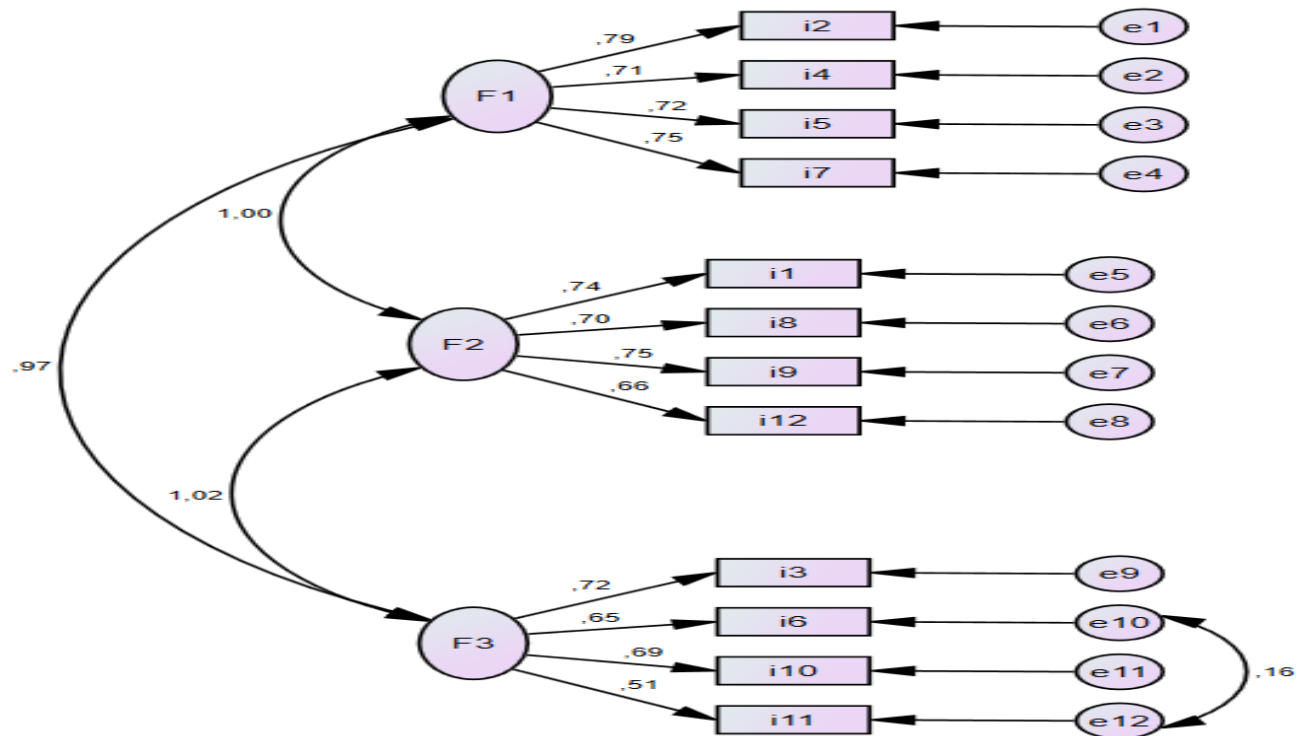
Scale and Subscales	F1	F2	F3	VAS - Peer Evaluation	Number of items	α	Ω	Results for the lower 27% and upper 27% groups	
								t*	p
F1	1	.		.572**	4	.832	.831	-30.496	.001
F2	.815**	1	.	.552**	4	.808	.808	-30.751	.001
F3	.737**	.784**	1	.520**	4	.757	.758	-28.443	.001
Scale Total	.927**	.940**	.903**	.594**	12	.921	.921	-37.541	.001

P: Independent Groups t test results **Correlation is significant at the 0.01 level (2-tailed); F: Factor; α : Cronbach's Alpha Value; Ω : McDonald's Omega Value

CMIN/DF=4.617, $\chi^2 p=0.001$, GFI=0.936, AGFI=0.900, CFI=0.950, RMSEA=0.078, RMR=0.019, SRMR=0.0382, NFI=0.937, TLI=0.933, IFI=0.9950, PGFI=0.600, PNFI=0.710

Figure 1.

PATH diagram and model fit indices related to the structure of the scale



psychometrically sound instrument specifically designed for peer evaluation in TBL contexts. Given the increasing adoption of TBL in health education programs in Türkiye, the availability of a validated Turkish-language scale enables standardized and comparable assessment of peer evaluation processes across institutions. Furthermore, by capturing core dimensions of TBL such as responsibility, initiative, and cooperation, the scale offers practical value for educators seeking to monitor group dynamics and provide formative feedback. From a measurement and evaluation perspective, the scale supports systematic, evidence-based assessment practices and facilitates both educational research and routine instructional evaluation. Study outcomes were examined in light of the existing body of literature.

Content Validity

In research involving the adaptation of scales into Turkish, it is crucial that the initial translation is performed by qualified English language experts, followed by a thorough assessment conducted by specialists in the relevant field (Seçer, 2020). In line with this approach, the scale was first translated by two professional English language translators and subsequently evaluated by a panel consisting of 12

subject matter experts. The CVI for individual items ranged from 0.80 to 1.00, leading to the retention of all items. A CVI score of 0.80 or higher is considered evidence that the items appropriately reflect the construct being measured and indicate strong agreement among the experts. In this study, the CVI values falling within the 0.80-1.00 range confirm that the items sufficiently represent the intended content (DeVellis & Thorpe, 2021).

Structure Validity

In scale adaptation studies, CFA is recommended for verifying a previously established structure (DeVellis & Thorpe, 2021; Seçer, 2020). In order to confirm the structural validity of the scale, CFA was performed. Prior to conducting CFA, however, it was necessary to verify that the data met the required assumptions and that the sample size was sufficient. Given the team-based learning context, each student provided multiple peer evaluations, which may partially weaken the assumption of independent observations. This data structure could potentially influence parameter estimates in confirmatory factor analysis and reliability analyses. However, the primary aim of the study was to examine the psychometric properties of the scale rather than to make individual-level inferences.

Treating each peer evaluation form as a unit of analysis is consistent with previous peer assessment and scale adaptation studies conducted in collaborative learning environments (Kweon & Park, 2023). Nevertheless, future studies may consider multilevel or clustered analytic approaches to further address data dependency.

The data structure involved repeated peer evaluations, resulting in 590 observations derived from 169 participants. This repeated-measures structure may partially violate the assumption of independent observations underlying CFA and internal consistency analyses. Consequently, model fit indices and reliability estimates may be somewhat inflated. However, the repeated evaluation process is inherent to team-based learning contexts and reflects the naturalistic use of peer evaluation instruments. Future studies may consider alternative analytical approaches, such as multilevel CFA or aggregation at the participant level, to further account for data dependency.

The suitability of the data for factor analysis was assessed through the KMO statistic, Bartlett's Test of Sphericity, and the anti-image correlation matrix. The literature suggests that a KMO coefficient greater than 0.60, anti-image values above 0.50, and a significant Bartlett's Test result serve as indicators of both sample adequacy and data appropriateness for conducting factor analysis (Pallant, 2020; Seğer, 2020). In the current study, the KMO value was calculated as 0.949, Bartlett's test yielded a significant result ($\chi^2=3624.740$, $p=.001$), and anti-image correlations ranged from 0.929 to 0.965. These findings demonstrate that the data meet the necessary conditions for factor analysis and that the sample size is adequate (Pallant, 2020). Furthermore, these findings are consistent with those reported by Kweon and Park, who found a KMO value of 0.922 and a significant Bartlett's test result ($\chi^2=2398.860$, $p<.001$) in their own study (Kweon & Park, 2023).

The congruence of the fit indices obtained from the CFA with the accepted value ranges in the literature corroborates the existing structure (DeVellis & Thorpe, 2021). In this study, the fit index values were CMIN/DF=4.617, χ^2 $p=.001$, GFI=0.936, AGFI=0.900, CFI=0.950, RMSEA=0.078, RMR=0.019, SRMR=0.0382, NFI=0.937, TLI=0.933, IFI=0.9950, PGFI=0.600, PNFI=0.710. These values align with those reported in the literature (Aksu et al., 2017; Bae, 2017; Kline, 2014). Kweon and Park reported that the fit index values of the 3-factor scale were 'CMIN/DF=2.345, χ^2 $p=.001$, CFI=0.970, RMSEA=0.060, SRMR=0.041, NFI=0.949, and TLI=0.957' (Kweon & Park, 2023). The fit indices obtained in this study and those reported by Kweon and Park demonstrate similarity and fall within the value ranges accepted in the literature (DeVellis & Thorpe, 2021).

The RMSEA value (0.078) should be interpreted in light of the relatively large number of observations ($N = 590$). In large samples, even minor model misspecifications may be amplified and reflected in absolute fit indices such as RMSEA. In addition, RMSEA is sensitive to model degrees of freedom and data structure, particularly in models with a limited number of indicators per factor and repeated measurement designs. Therefore, the observed RMSEA value likely reflects these methodological characteristics rather than substantive model misspecification. The adequacy of the model is further supported by incremental fit indices (CFI, TLI, IFI) and low residual-based indices (SRMR, RMR). Additionally, this result may be partially attributed to the relatively complex data structure arising from repeated peer evaluations, which may inflate residual covariance and affect absolute fit indices such as RMSEA.

The factor loadings obtained from the analysis ranged between 0.51 and 0.79. According to established guidelines in the literature, factor loadings should exceed 0.30 to be considered acceptable indicators of construct validity. In this study, the lowest factor loading observed was 0.51, which suggests that all items have a strong contribution to the underlying construct, thereby supporting the structural soundness of the scale (Seğer, 2020). These findings substantiated the structure of the scale. In addition, positive correlations were observed between the total scale and its subdimensions and the VAS for Peer Evaluation. This further supports both the internal consistency and the criterion-related validity of the scale (Seğer, 2020; Tavşancıl, 2018).

Additionally, the relatively high correlations observed among the subdimensions ($r = 0.73-0.81$) may suggest the presence of a strong overarching construct underlying peer evaluation in team-based learning. This raises the possibility that a unidimensional model could also be considered. However, the three-factor structure was retained based on theoretical considerations and the original scale structure, which conceptualizes peer evaluation as comprising responsibility, initiative, and cooperation (Kweon and Park, 2023). Moreover, the confirmatory factor analysis demonstrated acceptable model fit for the three-factor model, supporting the multidimensional representation of the construct. Future studies may further examine alternative models, such as a single-factor or higher-order factor structure, to better understand the dimensionality of the scale.

Reliability Analysis

Cronbach's Alpha value is the most commonly utilized method to ensure reliability in scale adaptation and development studies. Additionally, the McDonald's Omega value, parallel form method, and lower upper 27% quartile

analysis are alternative methods employed (DeVellis, 2016). Reliability coefficients such as Cronbach's Alpha and McDonald's Omega exceeding 0.60 are generally accepted as indicators of acceptable internal consistency, while values above 0.90 are considered to reflect excellent reliability. The very high Cronbach's Alpha value obtained in the pilot study ($\alpha = 0.965$) may be partially attributed to the data structure of the team-based learning context, in which participants provided multiple peer evaluations. Such repeated measurements may reduce error variance and artificially inflate internal consistency estimates. Therefore, the pilot reliability findings should be interpreted with caution. In the main study, reliability coefficients were lower yet still within acceptable ranges, supporting the internal consistency of the scale under more robust psychometric conditions.

In the present study, the "Responsibility" subdimension yielded a Cronbach's Alpha of 0.832 and a McDonald's Omega of 0.831. For the "Initiative" subdimension, both coefficients were calculated as 0.808. The "Cooperation" subdimension demonstrated reliability with a Cronbach's Alpha of 0.757 and a McDonald's Omega of 0.758. At the scale level, both Cronbach's Alpha and McDonald's Omega were found to be 0.921. These results collectively demonstrate that the scale and its subdimensions possess a high level of internal consistency and are statistically reliable (Pallant, 2020).

Kweon and Park (2023) reported that the Cronbach's Alpha coefficients for the "Responsibility," "Initiative," and "Cooperation" subdimensions were 0.840, 0.805, and 0.791, respectively. Similarly, the McDonald's Omega values for these subdimensions were reported as 0.840 for "Responsibility", 0.805 for "Initiative," and 0.802 for "Cooperation". The reliability coefficients obtained in the present study closely align with those reported by Kweon and Park, reinforcing the reliability of the adapted scale.

To evaluate the scale's discriminatory capability, mean scores of participants in the lowest and highest 27% groups were compared. A significant difference between these groups suggests that the items effectively distinguish between individuals with differing levels of the measured trait (Seçer, 2020). In the current study, significant differences ($p < .05$) were found between the lower and upper 27% groups for both the total scale and its subdimensions. These findings demonstrate that the scale effectively differentiates between participants with lower and higher levels of the measured traits, thereby confirming its adequate discriminatory power. In conjunction with the reliability analyses, these results provide further evidence of the scale's robustness.

Limitations and Strengths

Several limitations of this study should be acknowledged. First, the data structure involved repeated peer evaluations, resulting in multiple observations derived from the same participants, which may have partially affected the assumption of independent observations underlying the confirmatory factor analysis and reliability estimates. Although this structure reflects the inherent nature of team-based learning environments, it may have led to somewhat inflated model fit and reliability coefficients. Second, the sample distribution was not fully balanced in terms of gender and program representation, as participants were drawn from specific health programs, which may limit the generalizability of the findings to broader student populations. The high item means suggest a potential ceiling effect, which may limit the scale's sensitivity in distinguishing higher levels of peer evaluation performance. Finally, the relatively short duration of the team-based learning implementation may have constrained participants' familiarity with peer evaluation processes, potentially influencing response patterns. Despite these limitations, the study has several notable strengths, including a robust sample size in terms of total observations, the use of multiple validity and reliability indicators, and strict adherence to established scale adaptation guidelines. Moreover, the preservation of the original factor structure and the demonstration of acceptable model fit support the psychometric robustness of the Turkish version of the scale.

Conclusion and Recommendations

TBL is notably effective not only in fostering leadership, interpersonal communication, decision-making, and coordination skills during the learning process but also in enhancing students' academic achievement and addressing common challenges faced by team members. Additionally, TBL encourages students' motivation to learn and strengthens their sense of responsibility. In the literature, the measurement instruments that assess students' peer evaluations of TBL are limited. In national literature, there is no measurement instrument that evaluates students' peer assessments of TBL. Consequently, it is posited that a novel measurement instrument could inform future research.

Future research is recommended to further examine the psychometric properties of the existing Turkish version of the Peer Evaluation Scale for Team-Based Learning in diverse educational contexts. Specifically, studies using multi-center samples from different universities and academic disciplines are needed to evaluate the generalizability of the scale. In addition, test-retest

reliability analyses are recommended to assess the temporal stability of the Turkish form. Cross-cultural studies comparing the Turkish version with versions used in other cultural contexts may also provide valuable evidence regarding measurement invariance, without implying a re-adaptation of the scale. Such investigations will contribute to the broader validation and robust application of the scale in both educational research and practice.

The “Peer Evaluation Scale for TBL” comprises 12 items and 3 dimensions. The scale does not include a cut-off point or reverse items. With a Cronbach’s Alpha value of 0.921 and a McDonald’s Omega value of 0.921, the scale demonstrates high reliability. The scale’s conciseness, consisting of 12 items, enhances its utility. The “Peer Evaluation Scale for TBL” is a valid and reliable measurement instrument for determining students’ peer evaluation levels regarding TBL. The application of this scale in studies aimed at assessing students’ peer evaluation levels towards TBL is likely to further increase student engagement in the classroom.

Ethics Committee Approval: Ethics committee approval was received for this study from the ethics committee of Muş Alparslan University (Date: February 6, 2024, Number: 129968).

Informed Consent: Informed consent was obtained from all subjects involved in the study.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept-AK; Design-AK, GNY; Supervision-AK; Resources-AK, GNY; Data Collection and/or Processing-AK, GNY; Analysis and/or Interpretation-GNY; Literature Search-AK, GNY; Writing Manuscript-AK, GNY; Critical Review-AK

Conflict of Interest: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Use of Artificial Intelligence: In this study, an AI-assisted language editing tool (Paperpal) was used to improve the quality and accuracy of the English language. The tool was employed solely for language and grammatical editing, and the scientific content and academic responsibility of the manuscript remain entirely with the authors.

References

- Aksu, G., Eser, M., & Güzeller, C. (2017). *Exploratory and Confirmatory Factor Analysis and Structural Equation Model Applications*. Detail Publishing.
- Alberti, S., Motta, P., Ferri, P., & Bonetti, L. (2021). The effectiveness of team-based learning in nursing education: a systematic review. *Nurse Education Today*, 97, 104721. doi:https://doi.org/10.1016/j.nedt.2020.104721
- Bae, B.-R. (2017). *Structural equation modeling with Amos 24*. Chenngam Books, 76-309.
- Burgess, A., van Diggele, C., Roberts, C., & Mellis, C. (2020). Feedback in the clinical setting. *BMC Medical Education*, 20, 1-5. doi:https://doi.org/10.1186/s12909-020-02280-5
- Considine, J., Berry, D., Allen, J., Hewitt, N., Oldland, E., Sprogis, S. K., & Currey, J. (2021). Specific elements of team-based learning used in nursing education: a scoping review. *Nurse Educator*, 46(5), E84-E89. doi:10.1097/NNE.0000000000001061
- Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5(4), 194-197.
- DeVellis, R. F. (2016). *Scale development: Theory and applications*. 26: Sage publications.
- DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.
- Erkorkmaz, Ü., Etikan, İ., Demir, O., Özdamar, K., & Sanisoğlu, S. Y. (2013). Doğrulamalı faktör analizi ve uyum indeksleri. *Türkiye Klinikleri Journal of Medical Sciences*, 33(1), 210-223. doi:10.5336/medsci.2011-26747
- Goedde, R., & Sibley, J. (2014). *Approaches to peer evaluation: pro’s and con’s of various methods*. http://learntbl.org/wp-content/uploads/2014/06/Poster_TBL_peer_Feb2011-22nd.pdf
- Kline, P. (2014). *An easy guide to factor analysis*. Routledge.
- Kweon, Y.-R., & Park, J. (2023). Using the design-thinking method to develop and validate a peer evaluation scale for team-based learning (PES-TBL) for nursing students. *Nurse Education Today*, 127, 105849. doi:https://doi.org/10.1016/j.nedt.2023.105849
- Levine, R. E., & Hudes, P. D. (2021). *How-to guide for team-based learning*. Springer.
- Michaelsen, L. K., Davidson, N., & Major, C. H. (2024). Team-based learning practices and principles in comparison with cooperative learning and problem-based learning. *Journal on Excellence in College Teaching*, 3-4, 57-84.
- Michaelsen, L. K., & Sweet, M. (2011). Team-based learning. In L. K. Michaelsen & M. Sweet (Eds.), *New directions for teaching and learning* 1, 41-51.
- Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (3 ed.). McGraw-hill education (UK).
- Parmelee, D. X., & Michaelsen, L. K. (2010). Twelve tips for doing effective team-based learning (TBL). *Medical Teacher*, 32(2), 118-122. doi:https://doi.org/10.3109/01421590903548562
- Rezaee, R., & Moadeb, N. (2016). Team-based learning: A new approach toward improving education. *Acta Medica Iranica*, 54, 678-682.
- Rotgans, J. I., Schmidt, H. G., Rajalingam, P., Hao, J. W. Y., Canning, C. A., Ferenczi, M. A., & Low-Beer, N. (2018). How cognitive engagement fluctuates during a team-based learning session and how it predicts academic achievement. *Advances in Health Sciences Education*, 23, 339-351. doi:https://doi.org/10.1007/s10459-017-

9801-2

- Seçer, İ. (2020). *Psychological test development and adaptation process: SPSS and LISREL applications*. Ani Publications.
- Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9(1), 4-11. doi:10.12691/ajams-9-1-2
- Sisk, R. J. (2011). Team-based learning: systematic research review. *Journal of Nursing Education*, 50(12), 665-669. doi:https://doi.org/10.3928/01484834-20111017-01
- Tavşancıl, E. (2018). *Tutumların ölçülmesi ve SPSS ile veri analizi*. 6, Nobel Akademik Yayıncılık.
- Whitley, H. P., Bell, E., Eng, M., Fuentes, D. G., Helms, K. L., Maki, E. D., & Vyas, D. (2015). Practical team-based learning from planning to implementation. *American Journal of Pharmaceutical Education*, 79(10), 149. doi:https://doi.org/10.5688/ajpe7910149
- Woo, J. P. (2017). *The concept and understanding of structural equation model*. Hannarae.
- Zwarenstein, M., Goldman, J., & Reeves, S. (2009). Interprofessional collaboration: effects of practice based interventions on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews*, 3, CD000072. doi:https://doi.org/10.1002/14651858.CD000072.pub2