Behavior Intervention Competency Scale: Validity and Reliability Study

Özge Çevik – Necmettin Erbakan University, Türkiye 0000-0002-3633-370X
Hakan Sarı – Necmettin Erbakan University, Türkiye 0000-0003-4528-8936

Recepción: 04.07.2026 | Aceptado: 08.07.2026
Correspondencia a través de ORCID: Özge Çevik - 0000-0002-3633-370X
Citar: Çevik, Ö, & Sarı, H (2026). Behavior Intervention Competency Scale: Validity and Reliability Study. *REIDOCREA, 15*(46), 628-641.
Estudio de investigación: Este estudio se deriva de una parte de la tesis doctoral de la primera autora.
Área o categoría del conocimiento: Educación especial

Abstract: This study developed the Behavior Intervention Competency Scale to assess the behavioral intervention competencies of middle school teachers working with students with autism spectrum disorder. The study included 350 teachers from 10 subject areas. An initial 40-item pool was reviewed by three special education experts and one measurement and evaluation expert; 10 items were removed, resulting in a 30-item candidate form. Exploratory factor analysis was conducted with an independent sample of 185 teachers, and confirmatory factor analysis was performed with a separate sample of 165 teachers. Nine items were removed following exploratory analysis, and the remaining 21 items formed three subdimensions: Before Behavioral Intervention, During Behavioral Intervention, and After Behavioral Intervention. Confirmatory factor analysis supported the correlated three-factor model. Cronbach's alpha was .96 for the total scale and .86, .93, and .93 for the subscales. Composite reliability values ranged from .86 to .93, and average variance extracted values ranged from .53 to .67, supporting convergent validity. Discriminant validity was largely supported, although overlap was observed between the before- and during-intervention dimensions. A four-week test–retest analysis yielded a significant coefficient of r=.75. Overall, the scale demonstrated strong reliability and acceptable validity evidence.

Keywords: Behavioral Intervention

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized primarily by persistent difficulties in social communication and social interaction, alongside restricted and repetitive patterns of behavior, interests, or activities (American Psychiatric Association, 2022).

As the participation of students with ASD in general education settings becomes increasingly visible in the school years, schools are expected to address not only instructional adaptations but also systematic, preventive, and evidence-based approaches to classroom behavior support.

In Türkiye, inclusive education practices and special education services are implemented within the regulatory framework of the Ministry of National Education (MoNE). The Special Education Services Regulation defines the procedures and principles for ensuring that individuals with special education needs can access educational rights and services; it also frames inclusive education (mainstreaming/integration) as an approach in which students receive education alongside peers with appropriate supports (Ministry of Education, 2018). In addition, the MoNE Directorate General for Special Education and Guidance Services has issued updated guidance on inclusive education implementation to strengthen school-level processes (Ministry of National Education, Directorate General of Special Education and Guidance Services, 2017). Within this context, the quality of classroom-based behavior support is a key determinant of whether inclusive education goals translate into meaningful participation and learning opportunities for students with ASD.

A substantial portion of the challenges experienced by students with ASD in school settings is associated with challenging behaviors that may interfere with learning and social participation. When such behaviors are not effectively addressed, consequences can include reduced instructional time, weaker peer interactions, disrupted classroom climate, and increased teacher stress. Contemporary syntheses of intervention research indicate that multiple "focused" evidence-based practices can produce positive outcomes for learners with ASD when selected appropriately, implemented with fidelity, and monitored through data-based decision making (Hume *et al.*, 2021; Wong *et al.*, 2015). Accordingly, behavior intervention in schools should not be limited to "reducing problem behavior" but should be integrated with the teaching of functional skills, preventive environmental arrangements, and systematic monitoring.

At the core of evidence-based behavioral intervention is functional behavioral assessment (FBA), which aims to identify the functional variables maintaining behavior and to inform the design of function-based intervention plans. In secondary school contexts—where instructional demands, routines, and expectations diversify—teachers' capacity to conduct or participate in FBA processes becomes particularly salient (March & Horner, 2002). Importantly, the effectiveness of function-based interventions depends not only on plan quality but also on ongoing data collection, data review, and sustained implementation practices that support treatment fidelity (Pinkelman & Horner, 2017).

Beyond individualized supports, multi-tiered schoolwide frameworks also provide a preventive structure for strengthening behavioral climate. School-Wide Positive Behavioral Interventions and Supports (SWPBIS) is designed to promote shared expectations, explicit teaching of behavioral routines, reinforcement systems, and data-driven problem solving at the school level (Sugai & Horner, 2002). Group-randomized effectiveness research has reported that SWPBIS can yield positive impacts on organizational health and related student outcomes under implementation conditions aligned with the model (Bradshaw *et al.*, 2009). Such frameworks can also facilitate the alignment and sustainability of individualized behavior plans for students with ASD within broader school support systems.

Within these approaches, teachers' behavioral intervention competence emerges as a central implementation determinant. Behavioral intervention competence includes: (a) pre-intervention competencies such as operationally defining target behavior, collecting function-relevant data, and planning preventive arrangements; (b) during-intervention competencies such as selecting and implementing evidence-based strategies appropriately; and (c) post-intervention competencies such as monitoring outcomes, planning for generalization and maintenance, and updating intervention decisions based on data and stakeholder collaboration. Moreover, the linkage between intervention and student outcomes is closely tied to treatment integrity (implementation fidelity). Reviews in school-based behavior analytic research indicate that treatment integrity is meaningfully related to intervention outcomes and therefore should be assessed and supported in educational practice (Fiske *et al.*, 2008).

The proposed three-dimensional structure was specified a priori and was not derived solely from the empirical scree plot. It reflects the sequential but interrelated stages of a behavior-intervention process: pre-intervention competencies concerned with operational definition, assessment, and planning; during-intervention competencies concerned with implementation, monitoring, and data recording; and post-intervention competencies concerned with outcome evaluation, generalization, maintenance, and revision of intervention decisions. This process-based framework guided item generation, expert review, and subsequent factor labeling. Accordingly, the three factors were conceptualized as closely related domains nested within the broader construct of behavioral intervention competency.

Accurate assessment of teachers' competencies has strategic value for (i) conducting evidence-based needs assessments to design targeted in-service training and school-based professional development, and (ii) evaluating the effectiveness of implemented training programs. However, there remains an ongoing need for psychometrically robust instruments that can assess the behavioral intervention competencies of middle school teachers working with students with ASD through a process-oriented lens (before–during–after intervention) and in a manner aligned with school realities. Scale development literature emphasizes the importance of establishing content validity through expert review, evaluating construct validity through exploratory and confirmatory factor analytic procedures, and reporting reliability evidence (e.g., internal consistency, item–total relationships, discrimination indices) as core best practices (Boateng *et al.*, 2018; DeVellis, 2017).

Research questions

Accordingly, the purpose of the present study is to develop a valid and reliable scale to measure teachers' professional competencies in behavioral intervention. The resulting instrument is expected to enable comprehensive assessment across the stages of behavioral intervention (i.e., before, during, and after intervention), provide actionable evidence for needs-based professional development and in-service training planning, and contribute to the broader dissemination of effective behavior support practices in school settings. In the light of the purpose of the research, the research questions are shaped as follows:

- What is the factor structure (number of subdimensions and item distribution) of the scale developed to measure teachers' professional competencies in behavioral intervention?
- Is the validity of the scale (content and construct validity) supported by expert review and the results of EFA and CFA?
- Does the scale demonstrate acceptable reliability evidence, including internal consistency, item discrimination, subscale–total score relationships, and four-week test–retest stability?

Method

This section includes the preferred design, study group, data collection tools, activity implementation procedure and data analysis parts.

Research Design

This study is a scale development study conducted within a quantitative research framework. A measurement instrument was developed to assess teachers' professional competencies in behavioral intervention, and its psychometric properties were examined through exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Content validity was established via expert review, and reliability evidence was evaluated using internal consistency (Cronbach's Alpha), item–total correlations, and item discrimination analyses. Temporal stability was also examined through a four-week test–retest administration and the Pearson product–moment correlation between the total scores obtained at the two measurement occasions.

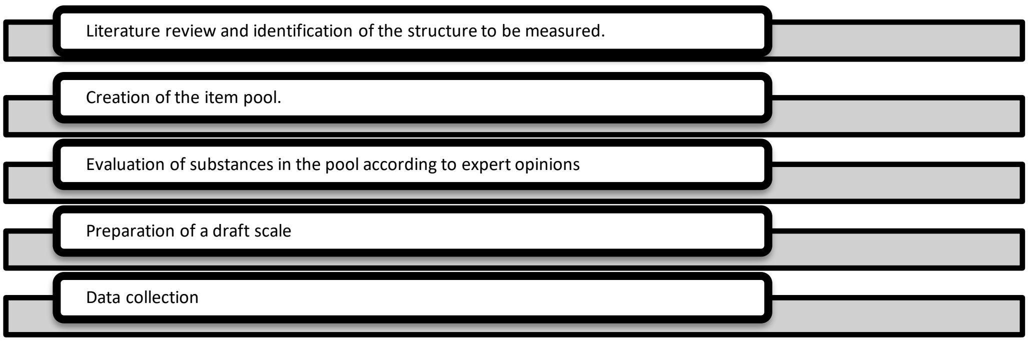
Study Group

The study group consisted of 350 teachers (EFA: 185, CFA: 165) working in middle schools affiliated with the Ministry of National Education who participated in the study on a voluntary basis. The participants represented 10 different subject areas. The study group was determined using a convenience sampling approach based on volunteer participation, in line with the aim of conducting the validity and reliability analyses required for scale development. Data obtained from the teachers were analyzed to examine the psychometric properties of the developed scale. The psychometric analyses were conducted on two completely independent samples. The factor structure was explored using data from the first sample (n = 185). The measurement model identified through EFA was then tested in a second sample comprising different participants who were not included in the first sample (n = 165). Thus, the structure developed using sample-specific information in the EFA was not retested on the same data, and the CFA provided independent cross-validation evidence. Conducting EFA and CFA on separate samples is consistent with recommended cross-validation practices in scale-development research (Lorenzo-Seva, 2022; Worthington & Whittaker, 2006).

The scale aims to measure teachers' professional competencies in identifying problem behaviors, defining target behaviors, selecting appropriate intervention strategies, and managing these processes. During the scale development process, opinions of field experts were obtained, content validity was established and validity and reliability analyses were conducted following the pilot implementation.

In this study, scale development was carried out in accordance with the procedural steps proposed by DeVellis (2017) and Tezbaşaran (1997). Accordingly, Figure 1 presents the procedural steps of the scale development process, and details of the development process are explained under the headings presented below.

Figure 1.
Scale development process steps.



Literature review and identification of the construct to be measured

In the first stage of the study, the literature was reviewed in detail to determine potential indicators related to the scale. In this context, national and international studies were examined and statements required for the scale were identified. Within the scope of the scale topic, different parameters were determined, and an item pool consisting of a total of 40 items was created.

Development of the item pool

The items considered appropriate for inclusion in the pool were reviewed by three experts in special education and one expert in measurement and evaluation. Each item in the pool for which expert opinions were obtained was evaluated under two headings: relevance to the scope of the construct and clarity/comprehensibility of the items. In addition, attention was paid to compliance with Turkish spelling and writing conventions.

Evaluation of the items in the pool based on expert opinions

Content validity was examined in line with the guidance of experts' judgments using the "content validity coefficient" developed by Veneziano and Hooper (1997). For each item, a coefficient value was obtained by calculating the ratio of the number of experts providing a favorable opinion to the total number of experts (minus one). Items with coefficients below .80 were removed. Accordingly, 10 items were deleted from the initial pool of 40 items.

Preparation of the draft scale

As a result, a 30-item form was prepared for the pilot implementation. A five-point Likert-type rating scale was used to determine participants' level of agreement with each item: "strongly disagree (1)," "disagree (2)," "neither agree nor disagree (3)," "agree (4)," and "strongly agree (5)." After the draft scale was prepared, it was administered as a pilot to 10 individuals. In particular, the clarity of the items and whether they fell within the intended scope were tested. Based on the obtained and expert opinions, it was concluded that the scale items and the overall instrument were functional, that the items were clear, and that they were within the scope of the construct.

Data collection

Following the pilot implementation, the 30-item candidate scale was transferred to Google Forms and converted into a digital format. The initial data-collection process lasted approximately four weeks. During the first two weeks, the candidate form was administered to 185 participants; during the subsequent two weeks, it was administered to a different group of 165 participants. The second group's data were stored separately and were not examined during the EFA, item-deletion decisions, or factor-retention process. The revision following EFA consisted solely of deleting nine items; the wording, response categories, and scoring of the retained items were not changed. CFA was subsequently conducted using the responses to the retained 21 items from the independent second sample. After screening, data from all 350 participants were suitable for their designated analyses (EFA: n = 185; CFA: n = 165). For the temporal-stability analysis, the scale was readministered to the same 350 participants four weeks after their initial administration under comparable conditions, and total scores based on the 21 retained items were calculated for both measurement occasions. The distributions of the two independent samples by age, gender, and educational level are presented in Table 1.

Table 1.
Descriptive characteristics of participants from whom data was collected

		EFA		CFA	
	Participant	f	%	f	%
Age	22-30	67	36	57	34.54
	31-39	84	46	73	44.24
	40 and above	34	18	35	21.22
Gender	Female	111	60	80	48.48
	Male	74	40	85	51.52
Graduation	Bachelor's degree	94	51	85	51.51
	Master's degree	88	47	75	45.46
	Doctorate	3	2	5	3.03
Total	Total	185	100	165	100

According to Table 1, sample for EFA the age range of the participants who took part in the scale development study varied from 22 to 40 years and above. While 67 participants (36%) were in the 22–30 age group, 84 participants (46%) were in the 31–39 age group, and 34 participants (18%) were aged 40 years and above. Of the participants, 111 (60%) were female and 74 (40%) were male. Regarding educational level, 94 participants (51%) held a bachelor's degree, 88 (47%) held a master's degree, and 3 (2%) held a doctoral degree. In addition, sample for CFA the age range of the participants who took part in the scale development study varied from 31 to 39 years and above. While 57 participants (34.54%) were in the 22–30 age group, 73 participants (44.24%) were in the 31–39 age group, and 35 participants (21.22%) were aged 40 years and above. Of the participants, 80 (48.48%) were female and 85 (51.52%) were male. Regarding educational level, 85 participants (51.52%) held a bachelor's degree, 75 (45.46%) held a master's degree, and 5 (3.03%) held a doctoral degree

Scale Development Procedure

In light of the organized data, validity and reliability analyses were conducted using two statistical programs. Before EFA and CFA, multivariate normality was examined using Mardia's multivariate skewness and kurtosis coefficients. The results indicated multivariate non-normality (Mardia's multivariate skewness = 325.59, χ^2 = 10039.13, $p < .001$; multivariate kurtosis = 1215.82, z = 39.70, $p < .001$). Therefore, Principal Axis Factoring (PAF), which is less dependent on multivariate normality than maximum-likelihood extraction, was used for EFA (Fabrigar et al., 1999). Promax rotation was selected because the dimensions were theoretically expected to correlate. Robust Maximum Likelihood (MLR) estimation was used for CFA to obtain robust parameter estimates and fit indices under multivariate non-normality (Finney & DiStefano, 2006). All EFA results reported in the manuscript—including the pattern coefficients, initial eigenvalues, and scree plot used as factor-retention diagnostics—were generated within the PAF procedure; principal component analysis was not used as the extraction method.

Sample adequacy was evaluated separately for each psychometric analysis rather than by using the total sample of 350 participants. For EFA, the 185 participants responding to the 30-item candidate form yielded a participant-to-item ratio of 6.17:1. For CFA, the independent sample of 165 participants and the retained 21-item model yielded a ratio of 7.86:1. These ratios were considered together with the magnitude of the communalities and factor loadings, the number of indicators per factor, the KMO statistic, model convergence, and global fit indices. This approach is consistent with evidence showing that factor-analytic sample requirements depend on model characteristics and indicator quality rather than on a single universal participant-to-variable rule (MacCallum et al., 1999; Mundfrom et al., 2005).

Through EFA, the factor loadings of the retained items, the relationships among the resulting subdimensions, and the explanatory power of the factors in accounting for the proposed model were examined. The accuracy of this structure was subsequently tested via CFA (Özdamar, 2004). In addition, CFA was used to evaluate item fit indices and the correlational relationships among the subdimensions. To examine whether the scale had a valid structure, EFA was conducted using statistical program and CFA was conducted using statistical program. Following construct validity analyses, the scale's internal consistency was tested through item discrimination tests and reliability analyses. In addition, temporal stability was evaluated by calculating the Pearson product–moment correlation between the total scores obtained from the two administrations conducted four weeks apart.

Results

Exploratory Factor Analysis (EFA)

To determine the suitability of the data for EFA, the results of the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's Test of Sphericity are presented in Table 2.

Table 2.
KMO and Bartlett's Test Results for the Behavioral Intervention Competency Scale

Kaiser–Meyer–Olkin (KMO) Value		.937
Bartlett's Test of Sphericity	$\chi 2$	5811.97
	df	210
	p	.00

In order to proceed with exploratory factor analysis (EFA), it is stated that the Kaiser–Meyer–Olkin (KMO) value—an essential prerequisite—should be above .60 (Pallant, 2005). In line with this criterion, the KMO value obtained in the present study was .94, which is above the threshold values recommended in the literature. In addition, Bartlett's Test of Sphericity was statistically significant, $\chi^2(210) = 5811.97$; $p < .001$. These findings indicate that the data were suitable for factor analysis. Following the calculation of the KMO value, the factor analysis was continued.

Factor retention was determined through a combined evaluation of the theoretical framework, initial eigenvalues, the scree plot, the interpretability of the pattern matrix, and simple-structure criteria rather than by any single statistical rule. Three initial eigenvalues exceeded 1.00. Items 4, 14, and 30 were first removed because they displayed salient cross-loadings with loading differences smaller than .10. After rerunning the PAF–Promax solution, Items 6, 9, 22, and 25, followed by Items 19 and 26, were removed for the same reason. The final solution contained no problematic cross-loadings and yielded three interpretable factors corresponding to competencies before, during, and after behavioral intervention. Promax rotation was retained throughout the analysis because the factors were expected to correlate.

Table 3.
Promax-Rotated Pattern Coefficients and Item Distribution

Item	Factors		
	Factor 1	Factor 2	Factor 3
1	.77		
2	.85		
3	.76		
5	.52		
7	.65		
8	.50		
10		.71	
11		.77	
12		.58	
13		.72	
15		.62	
16		.68	
17		.70	
18		.74	
20			.75
21			.69
23			.85
24			.84
27			.68
28			.64
29			.69

Factor loadings of the scale items are presented in Table 3. According to the analysis results, the scale yielded factors with eigenvalues greater than 1, and the factor loadings ranged from .50 (minimum) to .85 (maximum). This indicates that all items had factor loadings above .30. In the literature, a factor loading of at least .30 is considered the minimum criterion for an item to be regarded as meaningful (Tabachnick & Fidell, 2013). In addition, Hair *et al.* (2019) classify loadings of .40 and above as acceptable and .50 and above as high. Kline (2016) further emphasizes that factor loadings exceeding .50 provide strong evidence for construct validity. Therefore, the findings suggest that the scale items meet both the basic criteria and the higher standards reported in the literature. Following these calculations, the factor analysis was continued, and the total variance explained by the scale is presented in Table 4.

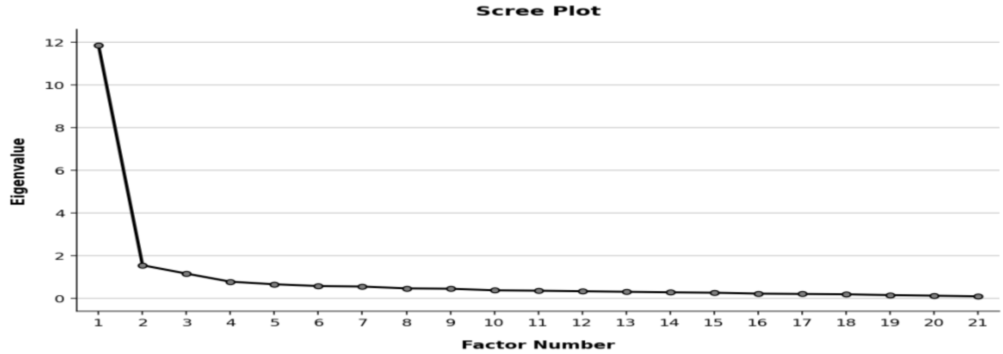
Table 4.
Initial Eigenvalues Used in Factor Retention Under PAF

Factor	Initial eigenvalue	% of total variance	Cumulative %
1	11.837	56.366	56.366
2	1.553	7.393	63.760
3	1.160	5.522	69.281
4	.783	3.730	73.011
5	.662	3.151	76.162
6	.580	2.763	78.926
7	.556	2.647	81.572
8	.470	2.237	83.809
9	.454	2.162	85.971
10	.379	1.806	87.777
11	.362	1.723	89.500
12	.336	1.598	91.099
13	.311	1.480	92.579
14	.285	1.356	93.935
15	.267	1.271	95.206
16	.225	1.072	96.277
17	.209	.995	97.272
18	.193	.921	98.193
19	.155	.740	98.933
20	.129	.613	99.546
21	.095	.454	100.000

Table 4 reports the initial eigenvalues of the 21 retained items used as factor-retention diagnostics within the PAF workflow. The first three eigenvalues were 11.837, 1.553, and 1.160, and together represented 69.28% of the total variance in the unreduced correlation matrix. These initial-eigenvalue percentages are reported to document the retention decision and should not be interpreted as results from a principal component extraction or as extraction-specific estimates of common variance.

Figure 2 presents the initial eigenvalues used as one component of the factor-retention decision within the PAF analysis.

Figure 2.
Scree Plot of Initial Eigenvalues Used in the PAF Retention Decision.

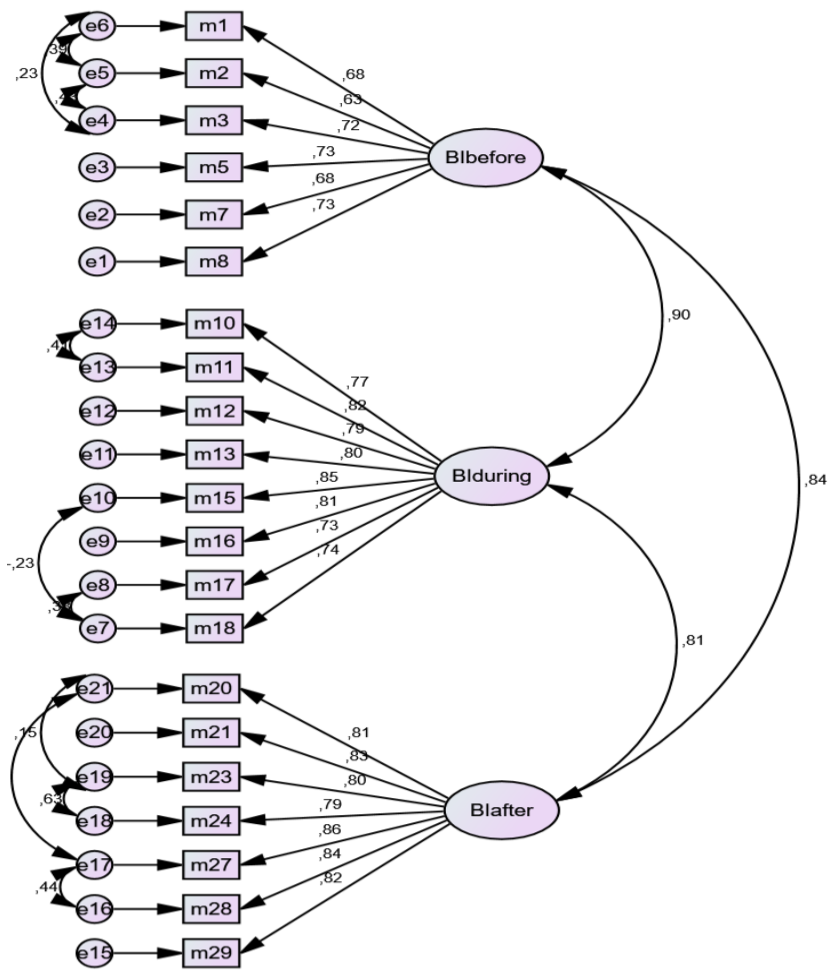


The scree plot displays a pronounced first eigenvalue, indicating a strong general behavioral-intervention competency component, while the curve continues to decline through the second and third eigenvalues before becoming comparatively flat. Because a scree plot can be open to interpretation, it was not used as a stand-alone basis for retaining three factors. The decision was made by jointly considering the three eigenvalues greater than 1.00, the stable and interpretable PAF–Promax pattern matrix, the presence of seven retained indicators per factor, the theoretically specified before–during–after framework, and the fit of the three-factor model in an independent CFA sample. Accordingly, the findings are interpreted as evidence for three strongly correlated and theoretically meaningful subdimensions within a broader competency construct; they do not imply that a unidimensional alternative has been definitively ruled out.

Confirmatory Factor Analysis (CFA)

To evaluate the structure identified through EFA, CFA was conducted on the second, non-overlapping sample (n = 165). The second sample had completed the same 30-item candidate form, but only responses to the 21 items retained through EFA were included in the CFA model. Because these data were not examined during item selection or factor retention, the CFA constituted an independent test of the three-factor measurement model shown in Figure 3.

Figure 3.
Standardized Three-Factor CFA Model of the Behavior Intervention Competency Scale.



When the fit indices of the model presented in Figure 3 were examined, the chi-square value (*χ²)* was found to be 403.701 with 163 degrees of freedom (df). The χ^2/df ratio calculated based on these values was 2.47. Çokluk *et al.* (2010) state that model fit is acceptable when this ratio is below 3. With respect to other fit indices, GFI (.91) and AGFI (.90) meeting the criterion of ≥ .90 indicate "good fit," as suggested by Hooper *et al.* (2008). Similarly, NFI (.90) and NNFI (.92) values also suggest good model fit at values of .90 and above, as reported by Tabachnick and Fidell (2007). The CFI value was .95, and Marsh *et al.* (2006) note that CFI values above .90 indicate strong model fit. The standardized RMR (SRMR) value was calculated as .01; Byrne (1994) emphasizes that SRMR values below .08 reflect good fit. Finally, the RMSEA value was .062, indicating an acceptable level of fit according to the recommended cutoff of ≤ .08 (Sarmento & Costa, 2019). Overall, some of the fit indices indicated good fit, whereas others reflected excellent fit. Following the initial confirmatory factor analysis (CFA), modification indices (MI) were examined to identify potential sources of localized strain within the model. The proposed modifications were evaluated by considering MI values together with the theoretical coherence and semantic similarity of the related items. Based on these evaluations, covariance paths were established between several item residuals within the same latent dimensions, including Items 1–2, Items 2–3, Items 15–16, Items 16–17, Items 20–21, Items 23–24, and Items 27–28. These modifications were considered theoretically justifiable because the paired items reflected conceptually overlapping aspects of behavioral intervention competencies and shared similar wording structures. Correlating the residuals of items within the same factor is consistent with recommendations in the structural equation modeling literature when supported by theoretical rationale and item content similarity (Byrne, 2016; Kline, 2016). Following these modifications, the fit indices demonstrated improved and acceptable model fit values, supporting the construct validity of the three-factor structure of the scale.

In addition to the fit indices and covariance estimates obtained from CFA, the correlational relationships among the subfactors were also examined (Table 5). In this context, the correlations among the subfactors in the model were positive and relatively high. In linear relationships between two variables, the correlation coefficient ranges from −1 to +1; when interpreting this coefficient, values close to −1 indicate a perfect negative relationship, whereas values close to +1 indicate a perfect positive relationship (Hunt, 1986).

Table 5.
Interfactor Correlation Coefficients

Factors	BIBefore	BIDuring	BIAfter
BIBefore	1.00	.77**	.72**
BIDuring	.77	1.00	.74**
BIAfter	.72	.74	1.00

Note: **Correlation significance level: $p < .01$

As shown in Table 5, the interfactor correlations were positive and statistically significant ($p < .01$), ranging from .72 to .77. These coefficients indicate substantial overlap among the sequential stages of behavioral intervention and are consistent with the conceptualization of the factors as related domains of a broader competency construct. To evaluate convergent and discriminant validity more directly, composite reliability (CR), AVE, and the Fornell–Larcker comparisons were also examined (Table 6).

Convergent and Discriminant Validity

Table 6.
Composite Reliability, Average Variance Extracted, and Fornell–Larcker Matrix

Factor	CR	AVE	BI-Before	BI-During	BI-After
BI-Before	.86	.53	.73	.77	.72
BI-During	.93	.63	.77	.79	.74
BI-After	.93	.67	.72	.74	.82

Note. CR = composite reliability; AVE = average variance extracted. Diagonal values in the correlation matrix are the square roots of AVE; off-diagonal values are interfactor correlations.

Average Variance Extracted (AVE) and Composite Reliability (CR) values were calculated for the subdimensions using the standardized factor loadings obtained from the final 21-item CFA model. For the evaluation of convergent validity, a CR value of at least .70 and an AVE value of at least .50 were adopted as the reference criteria. An AVE value of .50 or higher indicates that the relevant latent variable explains at least half of the variance in its indicators, whereas a CR value of .70 or higher indicates that the indicators consistently represent the corresponding construct (Bagozzi & Yi, 1988; Fornell & Larcker, 1981).

The analyses yielded CR = .86 and AVE = .53 for the Before Behavioral Intervention (BI-Before) dimension, CR = .93 and AVE = .63 for the During Behavioral Intervention (BI-During) dimension, and CR = .93 and AVE = .67 for the After Behavioral Intervention (BI-After) dimension. All CR values exceeded .70, indicating adequate composite reliability. Similarly, all AVE values met or exceeded .50, supporting convergent validity for the three subdimensions. However, because the AVE value for BI-Before (.53) was relatively close to the recommended threshold, the convergent-validity evidence for this dimension was considered acceptable but comparatively marginal.

Discriminant validity was evaluated using the Fornell–Larcker criterion, according to which the square root of the AVE for each subdimension should exceed its correlations with the other subdimensions (Fornell & Larcker, 1981). In conclusion, all subdimensions met the composite-reliability and convergent-validity criteria. Discriminant validity was largely supported, but it remained limited for the BI-Before and BI-During dimensions because of their high correlation. This overlap may reflect the fact that these dimensions represent theoretically related and sequential stages of the behavioral-intervention process. Future research should therefore examine the heterotrait–monotrait ratio of correlations (HTMT) to provide a more comprehensive assessment of discriminant validity (Henseler et al., 2015).

Item Discrimination and Reliability Analyses

Following the EFA and CFA, several analytical procedures were conducted to examine the internal consistency of the scale and the discriminative power of the items. The reliability of the scale was evaluated through statistical methods based on internal consistency. The reliability of the Behavior Intervention Competency Scale was examined both for the overall scale and separately for each subdimension. In this process, the overall reliability coefficient and the subscale-specific reliability coefficients were calculated, and the resulting values were reported comparatively. In addition, using the formulas presented below, reliability levels were determined for the total scale and for each factor.

In evaluating the internal consistency of scales, not only Cronbach's Alpha but also correlations between split halves and related reliability coefficients are taken into consideration. The literature notes that a Cronbach's Alpha coefficient above .70 is considered acceptable, whereas values of .80 and above indicate a high level of reliability (George & Mallery, 2019; Nunnally & Bernstein, 1994; Şencan, 2005). In addition, the split-half correlation, the Spearman–Brown formula, and the Guttman split-half coefficient were reported as supplementary indicators supporting the consistency of the scale. Obtaining coefficients above .80 in these analyses further confirms that both the overall scale and its subdimensions have strong internal consistency. Therefore, convergent findings obtained from multiple reliability methods provide scientific support for the Behavioral Intervention Scale as a reliable instrument for measurement purposes (Field, 2018; Tavşancıl, 2019). Moreover, based on the total factor scores, an independent-samples t test was conducted to examine the significance of the differences between the upper 27% and lower 27% groups.

The results of the t-test analysis indicate that the t values for the differences in factor scores ranged from 7.18 to 11.65. The differences between the factor scores of the upper and lower 27% groups were statistically significant for all factors at the .05 level ($p < .001$). In addition, the mean factor scores of the upper 27% group were higher than those of the lower 27% group

Test–Retest Administration

The test–retest method was used to determine the temporal stability of the Behavior Intervention Competency Scale (BICS). The scale was administered twice to the same participant group ($n = 350$) at a four-week interval. The interval was selected to be sufficiently long to reduce the likelihood that participants would remember their responses from the first administration, while remaining sufficiently short to minimize the likelihood of substantial change in the construct being measured. Both administrations were conducted under comparable conditions, and the relationship between the total scores obtained at the two measurement occasions was examined using the Pearson product–moment correlation coefficient.

Findings on Temporal Stability

To examine the temporal stability of the BICS, the total scores obtained from the two administrations conducted at a four-week interval were compared. The Pearson product–moment correlation analysis indicated a positive and statistically significant relationship between the total scores obtained from the first and second administrations, $r = .75$, $p < .01$. A test–retest coefficient of .70 or above is generally considered adequate evidence of score stability (Nunnally & Bernstein, 1994). Accordingly, the obtained coefficient indicates that the scale produced sufficiently stable and consistent measurements over the four-week interval and demonstrated adequate test–retest reliability.

Discussion

The construct validity of the Behavior Intervention Competency Scale was examined through EFA and CFA conducted on independent, non-overlapping samples. PAF with Promax rotation was used for EFA because the data were non-normal and the dimensions were theoretically related, whereas robust maximum likelihood estimation was used for CFA. The independent-sample design reduced the risk of capitalizing on sample-specific characteristics and provided cross-validation evidence for the proposed measurement model.

- As a result of the EFA, nine items that displayed problematic cross-loadings or failed the prespecified retention criteria were removed. The remaining 21 items formed three interpretable subdimensions corresponding to competencies before, during, and after behavioral intervention. The first eigenvalue was substantially larger than the subsequent eigenvalues, suggesting a strong general competency component; nevertheless, the three-factor solution was retained because it was theoretically specified, yielded a coherent simple structure, and was supported in the independent CFA sample.

- CFA supported the adequacy of the correlated three-factor model in the independent sample. However, the high interfactor correlations indicate that the subdimensions should be interpreted as closely related stages of a broader behavioral-intervention competency construct rather than as entirely independent traits.
- Cronbach's alpha was .96 for the total scale and .86, .93, and .93 for the three subdimensions. Composite reliability values (.86–.93) and AVE values (.53–.67) supported internal consistency and convergent validity. The Fornell–Larcker analysis provided discriminant-validity evidence for most factor pairs but indicated incomplete differentiation between the before- and during-intervention dimensions. The four-week test–retest coefficient ($r = .75$, $p < .01$) additionally supported the temporal stability of the total scale score.
- Taken together, the findings provide evidence that the scale can be used to assess behavioral-intervention competency as a broad construct represented by three closely related process stages. The evidence is strongest for reliability, convergent validity, and the overall three-factor model; claims regarding the distinctness of the before- and during-intervention subdimensions should remain cautious.
- Because the scree plot showed a dominant first factor, the interfactor correlations were high, and the Fornell–Larcker criterion was not fully satisfied for the before–during pair, future studies should compare the correlated three-factor model with theoretically plausible unidimensional, higher-order, and bifactor models in larger independent samples. Replication should also examine discriminant validity using additional criteria and evaluate whether reporting both total and subscale scores is empirically justified.

References

American Psychiatric Association (2022). Diagnostic and statistical manual of mental disorders (5th ed, text rev). American Psychiatric Association Publishing.

Bagozzi, RP, & Yi, Y (1988). On the evaluation of structural equation models. Journal of the Academy of Marketing Science, 16, 74–94. https://doi.org/10.1007/BF02723327

Boateng, GO, Neilands, TB, ..., & Young, SL (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers in Public Health, 6, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Bradshaw, CP, Koth, CW, Thornton, LA, & Leaf, PJ (2010). Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. Prevention Science, 11(2), 133–144. https://doi.org/10.1007/s11121-009-0119-1

Bryant, FB, & Yarnold, PR (1995). Principal components analysis and exploratory and confirmatory factor analysis. In LG Grimm, & PR Yarnold (Eds), Reading and understanding multivariate statistics (pp 99–136). American Psychological Association.

Büyüköztürk, Ş (2016). Handbook of data analysis for the social sciences (22nd ed.). Pegem Academy.

Byrne, BM (1994). Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming. SAGE Publications.

Cattell, RB (1978). The scientific use of factor analysis. Plenum Press.

Comrey, AL, & Lee, HB (1992). A first course in factor analysis (2nd ed.). Lawrence Erlbaum Associates.

Çokluk, Ö, Şekercioğlu, G, & Büyüköztürk, Ş (2010). Multivariate statistics for the social sciences: SPSS and LISREL applications. Pegem Academy.

DeVellis, RF (2017). Scale development: Theory and applications (4th ed.). SAGE Publications.

Everitt, BS (1975). Multivariate analysis: The need for data, and other problems. Heinemann Educational Books.

Fabrigar, LR, Wegener, DT, MacCallum, RC, & Strahan, EJ (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Field, A (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

Finney, SJ, & DiStefano, C (2006). Non-normal and categorical data in structural equation modeling. In GR Hancock, & RO Mueller (Eds), Structural equation modeling: A second course (pp 269–314). Information Age Publishing.

Fiske, KE, & Delmolino, KA (2008). Treatment integrity of school-based behavior analytic interventions: A review of the research. Behavior Analysis in Practice, 1(2), 19–25. https://doi.org/10.1007/BF03391724

Fornell, C, & Larcker, DF (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50. https://doi.org/10.1177/002224378101800104

George, D, & Mallery, P (2019). IBM SPSS statistics 26 step by step: A simple guide and reference (16th ed). Routledge. https://doi.org/10.4324/9780429056765

Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Lawrence Erlbaum Associates.

Hair, JF, Black, WC, Babin, BJ, & Anderson, RE (2019). Multivariate data analysis (8th ed.). Cengage.

Henseler, J, Ringle, CM, & Sarstedt, M (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43, 115–135. https://doi.org/10.1007/s11747-014-0403-8

Hooper, D, Coughlan, J, & Mullen, MR (2008). Structural equation modeling: Guidelines for determining model fit. Electronic Journal of Business Research Methods, 6(1), 53–60.

Hopkins, WG (2014). A new view of statistics. http://www.sportsci.org/resource/stats/effectmag.html

HumeK, Steinbrenner, JR, ..., & Savage, MN (2021). Evidence-based practices for children, youth, and young adults with autism: Third generation review. Journal of Autism and Developmental Disorders, 51(11), 4013–4032. https://doi.org/10.1007/s10803-020-04844-2

Hunt, RJ (1986). Percent agreement, Pearson's correlation, and kappa as measures of interexaminer reliability. Journal of Dental Research, 65(2), 128–130.

Kline, P (1979). Psychometrics and psychology. Academic Press.

Kline, P (1994). An easy guide to factor analysis. Routledge.

Kline, RB (2016). Principles and practice of structural equation modeling (4th ed.). Guilford Press.

Lorenzo-Seva, U (2022). SOLOMON: A method for splitting a sample into equivalent subsamples in factor analysis. Behavior Research Methods, 54(6), 2665–2677. https://doi.org/10.3758/s13428-021-01750-y

MacCallum, RC, Widaman, KF, Zhang, S, & Hong, S (1999). Sample size in factor analysis. Psychological Methods, 4(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

March, RE, & Horner, RH (2002). Feasibility and contributions of functional behavioral assessment in schools. Journal of Emotional and Behavioral Disorders, 10(3), 158–170. https://doi.org/10.1177/10634266020100030401

Marsh, HW, Hau, KT, ..., & Peschar, J (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural psychometric comparisons across 25 countries. International Journal of Testing, 6(4), 311–360. https://doi.org/10.1207/s15327574ijt0604_1

Ministry of National Education (2018). Regulation on special education services (Official Gazette No. 30471). https://mevzuat.meb.gov.tr/dosyalar/1963.pdf

Ministry of National Education, General Directorate of Special Education and Guidance Services (2017). Circular on inclusive/integrated education practices updated. https://orgm.meb.gov.tr/www/kaynastirmabutunlestirme-yoluyla-egitim-uygulamalari-genelgesi-guncellendi/icerik/934

Mundfrom, DJ, Shaw, DG, & Ke, TL (2005). Minimum sample size recommendations for conducting factor analyses. International Journal of Testing, 5(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Nunnally, JC, & Bernstein, IH (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Özdamar, K (2004). Statistical data analysis with software packages. Kaan Publishing.

Pallant, J (2005). SPSS survival manual: A step by step guide to data analysis using SPSS for Windows (Version 12) (2nd ed.). Open University Press.

Pinkelman, SE, & Horner, RH (2017). Improving implementation of function-based interventions: Self-monitoring, data collection, and data review. Journal of Positive Behavior Interventions, 19(4), 232–243. https://doi.org/10.1177/1098300716683634

Sarmento, RP, & Costa, V (2019). Confirmatory factor analysis—A case study. arXiv. https://arxiv.org/abs/1905.05598

Seçer, İ (2017). Psychological test development and adaptation process: SPSS and LISREL applications. Anı Publishing.

Sugai, G, & Horner, RH (2002). The evolution of discipline practices: School-wide positive behavior supports. Child & Family Behavior Therapy, 24(1–2), 23–50. https://doi.org/10.1300/J019v24n01_03

Tabachnick, BG, & Fidell, LS (2013). Using multivariate statistics (6th ed.). Pearson Education.

Tavşancıl, E (2019). Measurement of attitudes and data analysis with SPSS (6th ed.). Nobel Academic Publishing.

Veneziano, L, & Hooper, JA (1997). A method for quantifying content validity of health-related questionnaires. American Journal of Health Behavior, 21(1), 67–70.

Wong, C, Odom, SL, ..., & Schultz, TR (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. Journal of Autism and Developmental Disorders, 45(7), 1951–1966. https://doi.org/10.1007/s10803-014-2351-z.

Worthington, RL, & Whittaker, TA (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