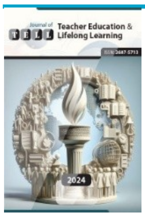

Journal of Teacher Education and Lifelong Learning (TELL)

Volume: 6 Issue: 1 Year: 2024 Research Article ISSN: 2687-5713

Turkish Adaptation of the Test of Narrative Language for Use in Preschools

Bayram Kara[1] **Ramazan Arı**[2]

[1] School of Foreign Languages, Selcuk University, Konya, Türkiye
bayramkara@selcuk.edu.tr

[2] Department of Phycology, Faculty of Social Sciences and Humanities, Konya Food and Agriculture University, Türkiye
ramazan.ari@gidatarim.edu.tr

Article Info

Article History
Received: 17/04/2024
Accepted: 03/06/2024
Published: 30/06/2024

Keywords:
Narrative,
Language,
Language development,
Validity,
Reliability

ABSTRACT

The purpose of this research is to adapt the Test of Narrative Language (TNL) to Turkish culture and to test its validity-reliability for Turkish children aged 60-72 months. The research was designed in general survey model. The sample included 240 five-year-old children attending a preschool in Konya city center. The TNL developed by Gillam and Peterson in 2004 was used to measure children's narrative skills within the scope of expressive and receptive language skills. The test includes six tasks in three formats (no picture, sequence pictures, and single picture) and two subtests (comprehension and narration). Content and construct validity analyses were performed to test the validity. Internal consistency, test-retest, split-half tests, and inter-rater consistency coefficients were calculated to determine its reliability. The results showed the Turkish adaptation of the TNL is a valid and reliable instrument to measure the narrative skills of 5-year-old Turkish children..

Citation: Kara, B. & Arı, R. (2024). Turkish adaptation of the test of narrative language for use in preschools. *Journal of Teacher Education and Lifelong Learning*, *6*(1), 186-206.

INTRODUCTION

Considered the golden age of development, the preschool period is a highly critical stage in terms of children's language development. Children's language skills therefore need to be supported through various activities and a rich stimulating environment. Accurate and valid assessment of the language in preschool children is essential for early identification of children with special educational needs, the design of the intervention and the evaluation of the outcomes of these interventions (Dale & Henderson, 1987). One important component of language development is the oral narration skills. The ability to tell stories, which is considered within the scope of narrative skills, plays a major role in the interpersonal communication as it is a part of routine life, social interactions, and academic activities, and is therefore seen as a very important skill for children's language development (Duinmeijer, de Jong, & Scheper, 2012). In typically developing children, this skill starts from preschool period and continues throughout the school years. Looking at the Preschool Education Program of the Ministry of National Education (MoNE) considering language development characteristics, it is seen that 48-60-month-old children can form sentences with 4-5 words, use conjunctions, answer questions such as why-how-who, respond to questions about short simple stories and talk about personal experiences. 60-72-month-old children can form sequential and compound sentences with six or more words, use past, present and future tenses, ask and answer questions such as "who, what, when, where, why and how", use conjunctions like "because, later", retell the stories just read aloud, and tell meaningful stories by establishing relationships between pictures, objects or events (MEB, 2013).

Narrative abilities are an important aspect and indicator of language development in young children. The concept of story involves a series of events, usually involving goal-oriented behaviors, sequenced according to time. Narration, in other words storytelling, is broadly defined as the oral/written presentation of events that are related in terms of cause and effect, or the oral/written transfer of a life experience in a specific time sequence (Peterson, 1990). Children begin to develop their language skills from the birth by interacting with other language users whose language skills have matured to a certain level, and by the ages of 3 to 4 years, they are capable of telling stories (Stadler & Ward, 2005). These narrative abilities then gradually develop over time. Typically developing children can comprehend and retell stories by the age of six (Merritt & Liles, 1987). Storytelling requires more complex language and a higher level of thinking than is required for everyday conversations. In order to describe an event to a listener who does not share it, the storyteller must choose explicit vocabulary, use clear pronoun references and descriptive language, and describe in a logical order the events that constitute a story (Petersen et al., 2010).

Children's narrative language abilities can be measured in a variety of ways and the most commonly used ones are personal story production and story retelling with or without picture cues. Personal story production usually involves asking the child to create/tell a story based on a picture or their life experiences. Because of its reliance on personal experiences and its common use in young children's everyday language, story generation is inherently a good indicator of the natural form of spoken language (Hudson & Shapiro, 1991). Story retelling, in the broadest sense, is the retelling of a story in the child's own words and expressions. The child listens to a story with or without picture support and is asked to retell the story. Morrow (1986) defines story retellings as post-listening or post-reading recollections in which the reader or listener recounts what they remember in oral or written form. Children's retelling of the story reveals not only what they remember but also how much they understand (Boudreau, 2008). Story retelling helps children to organize the various details of the story and to sequence the story events. Getting children to retell stories read by their parents or teachers is a widely used strategy that supports story comprehension and expressive vocabulary (Gambrell & Dromsky, 2000). In order to understand a story, children need to be aware of what is important in the story and use this knowledge to make the story understandable. Awareness of the elements that make up the story, i.e. the overall structure of the story, has a positive impact on the development of various literacy skills. Mandler and Johnson (1977) found that children of all ages use their knowledge of story structure to help them remember important details in a story. Knowledge of story structure matters when children interpreting and constructing their stories (Golden, 1984). As stated by Bower (1976), children who are not aware of story structure tell stories in which some story elements are missing, misordered, and lack harmony or coherence. Morrow (1986) concluded that when preschool children are encouraged to tell stories read to them, their comprehension and use of verbal language skills improve. Mages (2008) argues that children's ability to tell stories is related to academic literacy. Davies (2007) reported that storytelling improves children's listening and speaking skills and promotes language and imagination development.

In conclusion, storytelling and comprehension are seen as two important indicators of children's language development. It is important to evaluate children's storytelling skills and support them when necessary, as it forms the basis for their academic success. It helps them establish a healthy communication with the people around them, thus playing a major role in their social development. The TNL tests the narrative language abilities of children through stories. A problem that arises here is that there is no measurement tool in Turkey that can evaluate and measure children's expressive and receptive language on the basis of storytelling skills, both in terms of how well they understand the stories just read to them and how well they can create original or personal stories. Since the current measurement tools are based on measuring children's expressive and receptive language skills through concepts, words and phrases, the lack of a story-based measurement tool is considered a problem. In addition to assessing children's narrative abilities, the TNL has important areas of use such as identifying children with developmental language disorders, determining whether there is a discrepancy between the levels of receptive and expressive language development, and evaluating the effectiveness of an educational program implemented to support language development in children. Therefore, the results of this research are considered significant as they will increase the diversity of existing measurement tools and introduce a measurement tool that can be used by educators and researchers to evaluate children's oral language development.

The main purpose of this study is to develop a Turkish version of the TNL and establish its validity and reliability for 60-72-month-old Turkish children. In particular, the study seeks to address the following research questions:

1. Is the Turkish version of the TNL a valid assessment tool for 60-72-month-old children?

1.1. Does the Turkish version of the TNL meet the content validity criteria for 60-72-month-old children?

1.2. Does the Turkish version of the TNL meet the construct validity criteria for 60-72-month-old children?

2. Is the Turkish version of the TNL a reliable assessment tool for 60-72-month-old children?

2.1. Do the results of the internal consistency analysis calculated for the Turkish version of the TNL show that the test is a reliable tool?

2.2. Do the results of the test-retest reliability analysis calculated for the Turkish version of the TNL indicate that the test is a reliable tool?

2.3. Do the results of the Split-half reliability analysis calculated for the Turkish version of the TNL show that the test is a reliable tool?

2.4. Do the results of the interrater reliability analysis for the Turkish version of the TNL indicate that the test is a reliable tool?

METHOD

Research Design

This research was done to develop a Turkish version of the TNL and establish its validity and reliability for 60-72-month-old Turkish children. Given that, it was conducted based on a general survey design. Such designs attempt to describe a present or past situation in their existing condition (Karasar, 2013).

Participants

The sample of the study included 240 five-year-old children with normal development, attending preschools in Konya city-center. Simple random sampling method was used for the selection of the participants. In determining the sample size for the validity and reliability of the TNL, expert opinions stating that approximately three times the number of five-year-old children (n=83) included in the original version of the test would be sufficient were taken into consideration and the study was conducted with 240 children. In order to eliminate bias in the study, children were randomly selected and the sample eventually included typically developing preschoolers with parental consent. Those who did not want to participate in the study were left out and the data belonging to children who failed to complete the test and could not fulfil the instructions properly (n=5) were excluded from the analyses.

Table 1. *Distributions by age and gender*

Variable		60-66 Months		67-72 Months		Total	
		n	%	n	%	n	%
Age		127	53	113	47	240	100
		n	%	n	%		
	Girls	57	24	52	21.5	109	45.5
Gender	Boys	70	29	61	25.5	131	54.5
	Total	127	53	113	47	240	100

Research Instruments and Processes

The research data was collected using the Test of Narrative Language (TNL) developed by Gillam and Peterson (2004) to measure the narrative skills of children.

Test of Narrative Language (TNL)

The TNL (Gillam & Peterson, 2004) is an instrument developed to assess the narrative comprehension and production in children (n=1059) from the ages of 5 years-0 months to 11 years-11 months through three types of stories; a script, a personal narrative and a fictional narrative. The test consists of six tasks organized into three different formats (no pictures, sequenced pictures and single picture) and two subtests (comprehension and oral narration). The test is administered to children individually and the administration time for each child varies between 15 and 25 minutes depending on the child's performance.

Comprehension skills are measured by the child's responses to the questions posed after the oral presentation of the stories. The comprehension subtest includes three tasks. In the first task (Task No. 1), no picture support is provided. The child listens to a short story and then answers the questions asked by the examiner. In the second task (Task No. 3), the child is presented with 5 pictures appropriate to the flow of events in a story. The child is asked to look at the pictures while listening to the story. Afterwards, questions are asked to measure the extent to which the child understands the story. The child is allowed to look at the pictures while answering the questions. In the third task (Task No. 5), the child is presented with a single picture related to a story. Again, the child is asked to look at the picture while listening and answers the story questions asked by the examiner.

The questions consist of literal ones, which are answered by directly accessing the information in the text, and inferential ones, which require interpreting the information explicitly given in the text through existing knowledge and establishing new associations between the ideas in the text. The questions aim to measure children's ability to understand words and sentences, as well as their ability to make connections between the main ideas of the theme or topic addressed in the story. Children are asked questions about specific information presented in each of the stories (e.g. the name of the characters, the events and the problem in the story) and are given one point for each correct answer.

Task 1 is characterized by the fact that after the story is read aloud by the examiner, the child is asked to propose a solution for the complex situation presented ("What do you think they should do now?"). This question aims to obtain information about the child's ability to propose a coherent solution to the problem in the story. The comprehension subtests can be scored simultaneously or later by listening to the recordings.

The oral narration includes the tasks of retelling the story read aloud without picture support (Task 2), creating a personal story based on 5 pictures depicting the main events (Task 4), and creating a personal story by looking at a single picture presented (Task 6). The retelling task (Task 2) asks the child to retell the story presented in Task 1. The child's performance is measured by looking at how much of the basic information (e.g., the setting, names of characters, conjugations of verbs) that has been predetermined to be scored in the story and giving one point for each correct answer. In Tasks 4 and 6, performance is measured based on the story elements produced by the child. This includes both macrostructure (setting, characters, story elements - problem situation, actions and events, temporal relationship, cause and effect, closure, coherence and creativity) and microstructure (vocabulary and grammar - identification of objects, use of pronouns, conjugation of verbs, grammatical structure of sentences).

There is no time limit for the 6 tasks in the test, but the time required to administer the test ranges from 15 to 25 minutes. Questions in the story production tasks are scored from 0 to 2 (e.g., 0 = three or more grammatical errors; 1 = one or two grammatical errors; or 2 = no grammatical errors). Oral narration tests are not scored instantly, but are scored after the child's stories, which are recorded with a voice recorder, are converted into written format. When scoring, the child's exact words should be taken into account, not what the child means or what the examiner infers from the child's words. The recordings of children's responses need to be listened to over and over again until it is ensured that all items have been scored correctly.

Data Analysis

Before the data collection process, the manual was examined in detail by the researcher in case of any possible problems that might be encountered. After reviewing the model practice and scoring sections of the manual, the test was administered and scored by the researcher to three children. In this way, competence was gained in using the test in an error-free manner. The teachers of the children and the administrators of the kindergartens were interviewed to inform them about the purpose and procedure of the study and the materials to be used, and appropriate days and times for the testing were determined.

Before testing the children, the researcher was introduced to the class by the teacher and briefly informed the children about the activity to be carried out in order for the children to gain familiarity with the researcher. The test then was administered to each child individually in a well-lit, quiet and distraction-free room. In order to increase the motivation of the children and prepare them for the activity, short conversations of approximately 5 minutes were held with children before starting the test. The booklet with colorful pictures of the stories included in the test was positioned in a way that the child could easily see them. The instructions and questions were directed in a tone of voice that the child could easily hear, and each stage of the test, which included six tasks, was recorded with a voice recorder. The testing time varied between 15 and 25 minutes depending on the child's performance. At the end of the test, the child was taken back to the classroom and the testing procedure continued with another child. Since the test was conducted individually and each test lasted around 20 minutes on average, no scoring was done during the implementation to save time and do a better assessment.

Research data were collected from a total of 240 children from five kindergartens in Konya. Then these data in audio format were uploaded to the computer. Children's responses to the questions measuring how much they understood the stories in the test were scored by listening to the recordings directly. The stories they told as a part of story generation tasks were transcribed by listening to them at least two or three times, and the texts obtained were analyzed in detail and scored in accordance with the criteria specified in the manual.

The adaptation process of the TNL into Turkish started with the translation of the original stories and test items into Turkish by four experts working in the field of English Language Teaching. Then, a fifth expert, also working in the field of English Language Teaching, translated the original stories and test items back into English using the "back translation technique" and compared them with the original stories and test items. It was seen that there was a unity of expression and meaning between the Turkish and the original stories and test items. Eight academicians working in the field of child development and education were asked to evaluate the suitability of the test in terms of ambiguity, accuracy and the suitability of the story illustrations for five-year-old children and Turkish culture, and to make suggestions if necessary. Based on the expert opinions, it was accepted that the TNL had content validity. Then, the test was administered to 5-year-old children (n=15). With this small preliminary study, it was seen that the stories and test items in the test were properly understood by the children.

Validity and reliability analysis of the data was carried out using the SPSS 22.0 program. For the validity, the collected data were tested according to two validity criteria; content validity and construct validity. In order to ensure content validity, the evaluations and opinions of academicians working in the relevant field were requested. Exploratory factor analysis technique was used to test the construct validity of the test. The Kaiser-Meyer-Olkin (KMO) coefficient and Bartlett Sphericity Test were used to assess whether the data were suitable for factor analysis. Maximum Likelihood Confirmatory Factor Analyses were conducted to see whether the TNL had a fit index and eight different data fit indices were calculated: Chi-Square (X2), Degrees of Freedom (Sd), Ratio of Chi-Square to Degrees of Freedom (X2/Sd), Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Comparative Fit Index (CFI) and Non-Normed Fit Index (NNFI).

To determine the reliability of the TNL, the collected data were tested based on four reliability measures: internal consistency, test-retest, split-test reliability and inter-rater consistency. Cronbach's alpha coefficients were used for internal consistency. Test-retest correlation coefficients were examined using the Sperman-Brown formula. In testing the split-half and inter-rater consistency reliability, correlation coefficients for all tasks were calculated using the Spearman-Brown formula.

Ethic

The author(s) confirm(s) that ethical approval was obtained from Selçuk University (Approval Date: 29 /09 /2021, 2021/1587)

RESULTS

Content and Construct Validity for the Turkish Version of the TNL

Content validity shows whether the test items are appropriate for the purpose of measurement and whether they represent the area to be measured (Karasar, 2013). Seeking expert opinion is one of the most commonly used methods to test content validity in research (Özgüven, 2011). Accordingly, for the content validity of the current research, eight experts working in the field of child development and education were requested to evaluate the appropriateness of the Turkish version of the stories, test items, instructions and pictures for five-year-old children. Experts unanimously stated that the test items were suitable for the purpose and offered several suggestions. In addition, the test was finalized by making the necessary corrections in terms of language and expression and cultural differences, and the content validity of the test was established.

Construct validity indicates whether the items developed to evaluate a certain behavior can measure it or to what degree they can measure it accurately. Exploratory factor analysis was used to examine the construct validity of the test. The suitability of the data for exploratory factor analysis was tested with the Kaiser Meyer-Olkin (KMO) coefficient and Bartlett Sphericity Test. The KMO coefficient tests the suitability and adequacy of the sample size for factor analysis. When the KMO coefficient approaches 1, it means that the data are suited for analysis. The KMO coefficient value for the test was calculated as ,972 and the results of the Bartlett Sphericity (X2=4,654; p<.01) and Chi-Square tests were found significant. In line with the results, it was observed that the data were adequate for factor analysis. In order to reveal the structures called factors or components, factor analysis was performed based on Principal Component Analysis. After the analysis, it was found that there were 6 factors greater than 1. The total explained variance ratio was 71.961%. In the original version of the test, it was stated that there were two factors and six tasks related to these factors (Gillam & Pearson, 2004). Table 2 presents the results of the exploratory factor analysis of the test items.

Table 2. *Results of exploratory factor analysis*

Factor 1		**Factor 2**	**Factor 3**	**Factor 4**		**Factor 5**	**Factor 6**
2,853		2,426	1,817	1,596		1,460	1,419
Item no	Factor loadings after rotation	Item no	Factor loadings after rotation	Item no	Factor loadings after rotation	Item no	Factor loadings after rotation
23	,970	28	,883	92	,673	78	,471
26	,968	60	,883	86	,670	68	,470
64	,966	89	,881	13	,657	59	,470
27	,966	49	,875	35	,630	9	,462
58	,965	51	,868	3	,630	70	,460
31	,962	82	,853	72	,616	42	,458
17	,959	32	,845	2	,614	84	,456
56	,956	91	,836	69	,608	94	,454
25	,956	66	,836	79	,605	44	,452
15	,949	90	,832	85	,604	41	,451
21	,948	40	,828	55	,604	77	,449
37	,948	97	,812	74	,597	73	,444
12	,945	75	,810	52	,597	67	,433
53	,944	62	,809	83	,587	87	,430
57	,941	47	,789	71	,585	6	,426
50	,941	30	,788	96	,577	65	,420
24	,934	61	,787	38	,555	7	,384
33	,930	34	,744	80	,543	43	,382
36	,928	29	,736	88	,522	11	,358
19	,923	16	,729	8	,511	45	,345
18	,911	20	,720	93	,502	39	,339
48	,905	14	,709	46	,501	5	,311
63	,901	54	,700	95	,479	4	,307
81	,884	22	,686	10	,472	76	,260
						1	,251

The data in Table 2 show that the factor loads of the test items vary between .21 and .97. Büyüköztürk (2013) suggests that items with factor loading of .30 and higher discriminate individuals well, items between .20-.30 can be removed from the test if deemed necessary or the item should be adjusted, and items with factor loading lower than .20 should be removed from the test. Based on these results, it can be concluded that all items in the test have high discrimination. No items were removed from the test and the test consisted of 97 items in total.

Confirmatory factor analysis was conducted to verify the factorial structure of the test. In scale development, confirmatory factor analysis is performed to test the accuracy of the factor structures determined after exploratory factor analysis. "Maximum Likelihood Confirmatory Factor Analysis" was used to test the fit index of the Test and eight different data fit indices were calculated; Chi-Square (X2), Degree of Freedom (df), Ratio of Chi-Square to Degree of Freedom (X2/df), Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Comparative Fit Index (CFI), and Non-Normed Fit Index (NNFI) (Şimşek, 2007). Table 3 shows the results of the Confirmatory Factor Analysis of the Test.

Table 3. *Results of confirmatory factor analysis*

Fit Index	X^2	df	X^2/df	RMSEA	GFI	AGFI	CFI	NNFI
	1572,36	1021	1,540	0,021	0,96	0,89	0,97	0,97

As seen in Table 3, the results of the analyses for the eight data fit indices are as follows: Chi-Square (X2) value 1572.36, Degree of Freedom (df) 1021, Ratio of Chi-Square to Degree of Freedom (X2/df) 1.540, Root Mean Square Error of Approximation (RMSEA) 0.021, Goodness of Fit Index (GFI) 0. 96, Adjusted Goodness Fit Index (AGFI) 0.89, Comparative Fit Index (CFI) 0.97 and Non-Normed Fit Index (NNFI) 0.97.

Table 4. *Results of the t-test regarding age factor*

	n	X	ss	t	df	p
60-66 months	127	43.03	9.35	-18.740	198	.000
67-72 months	113	65.94	6.10			

Table 4 shows that the mean TNL scores of the children differed according to their age in months and the mean scores increased with age. The analysis revealed a statistically significant difference between the mean scores at the 0.05 level (t=-18.740, p<.05). It is seen that the mean scores of the children aged 67-72 months (X=65.94) are higher than the mean scores of the children aged 60-66 months (X=43.03).

Results Regarding Reliability

Büyüköztürk (2020), defines reliability as the consistency of the answers given to the items of a data collection tool and shows the degree of stability of the measurement results. A reliable measurement tool is expected to yield the same or similar results when administered repeatedly under the same conditions. Reliability is related to the degree to which the data collection tool accurately measures the characteristic it is intended to measure. For reliability in the study, the data were analyzed in terms of internal consistency, two-half test, test-retest and inter-rater reliability.

Internal Consistency

The Cronbach's Alpha formula was used to determine the internal consistency reliability of the test. This is because there is a triple scoring system in the test. Reliability coefficients related to Cronbach Alpha formula calculated for test tasks are given in Table 5.

Table 5. C*ronbach alpha reliability coefficients for test tasks*

Tasks	Cronbach's Alpha Coefficients	Total
Task 1	.76	.78
Task 2	.79	
Task 3	.73	
Task 4	.78	
Task 5	.78	
Task 6	.87	

As seen in Table 5, the Cronbach's Alpha coefficients for the tasks one to six of the test were calculated as follows; .76, .79, .73, .78, .78 and .87 respectively. Büyüköztürk (2020) suggests that a Cronbach Alpha coefficient of .70 or higher calculated for a psychological test can generally be considered sufficient for the reliability of test scores. The values varying between .73 and .87 and the coefficient of .78 for all items can be seen as an evidence that the test is reliable.

Test-retest Reliability

Test-retest reliability is explained by the correlation between the scores obtained by administering a test to the same group twice at certain intervals. It can be suggested that an average of four weeks between two administrations is generally appropriate (Büyüköztürk, 2020). The test was administered to 20 children four weeks later by the researcher for test-retest reliability, and the correlation coefficients for the tasks are presented in Table 6.

Table 6. *Test-retest correlation coefficients for the tasks*

Tasks	r
Task 1	.83
Task 2	.84
Task 3	.84
Task 4	.86
Task 5	.86
Task 6	.87

As shown in Table 6, the correlation coefficients for the tasks one through six of the test were calculated as .83, .84, .84, .84, .86, .86 and .87, respectively. Spearman Brown's formula was used in the calculation since the test-retest data were not normally distributed. The results confirm the test-retest reliability of the Turkish version of the test.

Split-half Test Reliability

The split-half test reliability is calculated using the Spearman Brown formula based on the relationship between the two halves of the test by dividing the test items into two equal halves as odd-even, first half-last half or randomly (Büyüköztürk, 2020). In order to calculate the split-half test reliability, the test items in each task were divided into two halves and the correlation coefficient was calculated for each task using the Spearman Brown formula. The reliability coefficients are given in Table 7.

Table 7. *Split-half test reliability coefficients for the test tasks*

Tasks	r
Task 1	.74
Task 2	.76
Task 3	.74
Task 4	.77
Task 5	.81
Task 6	.83

As can be seen in Table 5, the correlation coefficients for the tasks one to six were found as follows; .74, .76, .74, .77, .81 and .83 respectively. The results seem to ensure the split-half reliability of the Turkish version of the TNL.

Inter-rater Reliability

Raters are the people who score or evaluate a particular phenomenon. If the raters give similar scores in their measurements, the results are deemed reliable. Leahy et al. (1993) suggest that the degree of agreement between the raters should be at least .80 in the evaluations made using the measurement tool. In the study, an associate professor in the field of child development and education was accepted as the second rater and re-administered the test to 10 randomly selected children. The consistency in scoring between the researcher (first rater) and the second rater was analyzed using the

Spearman Brown formula for each task of the test. The results can be seen in Table 8.

Table 8. *Inter-rater reliability coefficients for the test tasks*

Tasks	r
Task 1	.82
Task 2	.84
Task 3	.88
Task 4	.87
Task 5	.88
Task 6	.90

The results in Table 8 indicate that all reliability coefficients were above the accepted level of .80. The data set remained consistent between the raters, confirming the inter-rater reliability of the test.

DISCUSSION, CONCLUSION, RECOMMENDATIONS

Test Structure

When we take a look at the structure of the TNL, we see that it contains some basic core elements that can shed light on the oral narrative skills of both typically developing children and children with developmental language disorders. The first point that needs to be emphasized is that the test structure includes tasks to access both the comprehension and the oral narration of the stories. In other words, it has a structure that measures children's receptive and expressive language skills together and can reveal possible inconsistencies between these language skills. This is because children need to develop receptive (comprehension) and expressive (expressive) language skills in order to become individuals with effective communication skills (McIntyre, 2005). Therefore, it can be used in combination with other formal and informal language assessment tools to provide data and perspectives for the diagnosis and intervention of language development problems in children. Another aspect is that the test can provide access to information about macro and micro structures in storytelling that require the use of cognitive and linguistic skills. Justice et al. (2006) suggest that in the evaluation of children's narratives through stories, it is important to consider and analyze these two structures together because they provide information about the language proficiency that children use in their narratives. Similarly, Meier (2020) reports that examining both micro and macro-structural components of language together contributes more to the complete understanding of young children's language development than evaluating one or the other alone.

In addition, the test includes a combination of tasks commonly used to measure children's spoken language skills. Three of the six tasks aim to measure children's receptive language skills, i.e. comprehension skills. One of these tasks measures children's comprehension of stories read aloud to them without pictures and the other two tasks measure their comprehension when stories are supported with a single or multiple pictures. As is known, the most frequently used assessment practice for this purpose is to ask children questions about a story they have listened to with or without picture support and to measure the extent to which they understand the story based on their responses. Indeed, researchers suggest that children's oral narratives can be used effectively to find out about children's comprehension of the story (Morrow, 1990; van Kraayenoord & Paris, 1996). In addition, it is seen as a type of language facilitation strategy that supports children's understanding and use of language structures and academic content in texts when adults direct questions about the story and ask children to answer them (Milburn et al., 2014).

In order to measure expressive language skills, the test includes narrative formats with and without pictures. Pictures are of great importance because they reinforce the text and present significant clues to ensure comprehension and support imagination (Snaith, 2007). In the picture format, apart from the task of creating a story based on sequential pictures, there are also tasks of creating a story based on a single picture. Mills (2015) argues that picture-assisted narratives tend to show less complex structures with shorter terms in general compared to non-picture narratives, which points to the necessity of including different formats and tasks in the assessment process when it comes to assessing children's narrative skills in an effective and healthy way. Children's oral narratives based on wordless picture books reveal children's language and thinking skills more comprehensively compared to standardized tests that lack the ability to measure communicative features or assess skill areas in a limited way (Heilmann, Miller, & Nockerts, 2010).

Considering the type and content of the stories in the TNL, we see that in addition to the characters and events that can be encountered in real life, there are also stories with fantastic elements and unusual events. This has advantages for children's language development process and it therefore would be useful to use both of these types when evaluating children's language skills. Gamble and Yates (2008) believe that children need access to different and varied texts to improve their knowledge and skills in language. Imaginative elements in fictional stories attract children's attention more, motivate them more about the story, stimulate their imagination, which in turn encourages them to understand the story better (Weisberg et al., 2015). In other words, supernatural events in stories are not something that children can see every day, so they can lead children to learn more by increasing their interest. On the other hand, it has been found that children are much more likely to transfer the information in stories with real-world events and characters to their own lives than in stories with unrealistic elements (Walker et al., 2012). Richert et al. (2009) reported that children are better able to comprehend the cause-effect relationships and solutions in stories and transfer them to real life more easily when the character of the story is based on real life rather than a fantasy. Based on this information, we can think that the TNL has an approach to measure children's story performances more accurately by using stories that include both unreal and real-life characters and events and by maintaining a balance between the two genres.

Turkish Translation and Cultural Adaptation

One of the main problems related to translation is the idioms and complex contexts specific to the culture in which the language was born. Since the literal translation of the items affects test validity and reliability, it is important to ensure the cultural adaptation of the test items as well as the linguistic adaptation and to carry out the translation by taking into account the differences between cultures (Seçer, 2015). When we look at the process of translation and cultural adaptation of the TNL into Turkish, the fact that there were fewer complex grammatical structures, idiomatic expressions and culture-specific items in the original form contributed to a healthier translation from the source language to the target language, as there was no loss of content, information and meaning.

It is recommended to make at least two translations of the tests from the original language (source language) to the target language (Beaton et al., 2001; Seçer, 2015). Based on this, the translation of the TNL from the source language, English, into the target language, Turkish, was carried out separately by four academics working in the field of English language education, and then the four Turkish forms produced were compared and synthesized by the translators to finalize the test. The translators were thoroughly informed about the content and purpose of the test and the population it would be administered to, as this increases the quality of the translation and the consistency of the language used in the translated document and the original form (Kalfoss, 2019). The original and translated versions of the test were then reviewed, ensuring equivalence of words (semantic equivalence), idioms and colloquialisms (idiomatic equivalence), and concepts (conceptual equivalence) between the two versions.

In the next stage, the test was back-translated from Turkish to English. Tyupa (2011) argues that back translation is one of the most popular methods for assessing translation quality in international and cross-cultural social research. Back translation, as the name suggests, is a process in which the translated text is re-translated into the source language by a translator who has not seen the original text. This process allows for a general review of the translation quality, as well as the identification of potential problems that may arise from poor quality translation and adaptation (Hambleton, 2017). If any discrepancies are found between the back-translated text and the original text, this is taken as an indicator of translation errors in the target language version. During the back-translation process, translators should have no prior knowledge of the test and should not see the source or another language version before or during the back-translation (Wild et al., 2005). This ensures a completely objective back translation. In light of this information, the final Turkish version of the test was back-translated into English by another academic working in the field of English language teaching who had no knowledge of the test. Then, the semantic equivalence between the source and target forms was evaluated, and possible confusions, ambiguities and errors that may arise from language structures were reviewed. As a result, the stories and test items in both the original and translated versions were found to have unity in terms of expressions and semantics.

Validity

Validity is the degree to which a measurement tool can accurately measure a trait or behavior that it aims to measure independently without confusing it with any other trait (Başol, 2019; Büyüköztürk, 2020). Erkuş (2003) defines validity as the degree to which a measurement tool serves the purpose for which it was developed. Validity is traditionally divided into three categories: content, criterion, and construct validity (Brown 1996; Crocker & Algina, 2000; Baykul 2021). The importance of utilizing different types of validity rather than a single one is emhasized when ensuring the validity of measurement tools (Demir, 2017). In this study, content and construct validity were conducted for the validity of the TNL; criterion-related validity could not be realized due to the lack of a parallel or similar measurement tool with a similar structure and content in the Turkish literature. Actually, Saad et al. (1999) suggest that three validity criteria - content, criterion, and construct validity - are generally used to provide validity support and that these three general methods generally overlap and that one or more of them may be appropriate to ensure validity depending on the situation.

Content validity is related to the extent to which the instrument fully assesses or measures the relevant construct (Karasar, 2009; Başol, 2019). Expert opinion is one of the most frequently used methods to analyze content validity (Uzunboylu & Özdamlı, 2011), to remove inappropriate items and content from the scale (Kapuscinski & Masters, 2010), and to make language and cultural adaptation if necessary. Expert opinion is the most common logical way to test content validity (Büyüköztürk, 2020). Within the framework of content validity, following the translation of the TNL into Turkish, the opinions of eight academics working in the field of child development and education were requested to determine whether the stories, pictures, instructions, and the assessment criteria in the test were appropriate for five-year-old children and Turkish culture.

In addition to content validity, one of the ways that can be used to verify the validity of a measurement tool is construct validity. In construct validity, the term "construct" refers to the psychological trait measured in the test (Demir, 2017). Construct validity is defined as the degree to which the measurement tool can accurately measure an abstract concept (factor) within the scope of the behavior it aims to measure (Büyüköztürk, 2020). It is stated that construct validity covers other approaches to validity and that construct validity is therefore related to the evaluation of validity types (Kline, 2000; Şencan, 2005).

Şencan (2005), Kan (2019), and Baykul (2021) underline that when demonstrating the construct validity of a test or scale, more than one technique and method should be used together and convincing evidence should be presented collectively. Merenda (2017) stated that the first step to be taken in ensuring the construct validity is to analyze the factor structure of the measurement tool and compare this factor structure with the one in the original form (p. 337). Demircioğlu (2015) reported that factor analysis is one of the two most frequently used methods to investigate and ensure the validity of the measurement tool. Factor analysis attempts to summarize how people respond to several items in an instrument in terms of a minimum number of underlying constructs or "factors" (Martin & Ford, 2018). It is a technique that aims to discover a small number of conceptually meaningful new dimensions or factors by aggregating a large number of interrelated variables (Çilingirtürk, 2011). In measurement tools with a high number of items, this technique is used to make the measurement tool simpler by reducing the complexity in order to analyze the results more easily. Factor analysis is handled in two dimensions as exploratory and confirmatory factor analysis. The purpose of performing exploratory factor analysis is to explain the measurement with a small number of factors by combining variables that measure the same structure or feature (Büyüköztürk, 2020). Through exploratory factor analysis, sub-dimensions of the scale can be obtained. Confirmatory factor analysis can be used to verify whether a previously developed measurement tool measures the relevant theoretical construct and it is stated that it would be appropriate to use confirmatory factor analysis when adapting a measurement tool developed abroad to Turkish (Başol, 2019; Seçer, 2015).

The first step in factor analysis is to determine whether the data are suitable for factor analysis. Kaiser - Meyer - Olkin (KMO) and Barlett Sphericity tests are used to determine whether the data structure is suitable for exploratory factor analysis (Çokluk et al., 2021). The KMO value calculated for the Turkish version of the TNL was found to be .972. The KMO value is within the range of 0-1 and a lower value indicates that the data is not suitable for factor analysis. Kaiser (1974, p. 35) reported that a value between 0.50 - 0.60 is "miserable", between 0.60 - 0.70 is "mediocre", between 0.70 - 0.80 is "middling", between 0.80 - 0.90 is "meritorious" and above 0.90 is "marvelous". Therefore, the KMO value we obtained (.972) can be interpreted as the level of suitability of the data set for conducting factor analysis is "marvelous". In addition, the results of Barlett Sphericity test show that the Chi-Square value (X2=4,654; p<.01) is significant. Tatlıdil (2002) noted that if the Barlett Sphericity test is found to be significant, factor analysis can be started. In line with the results obtained, it was concluded that the data were suitable for factor analysis, and exploratory factor analysis was performed.

Principal Component Analysis was carried out to reveal the structures called factors which the test evaluated and it was observed that there were six factors with factor loadings greater than 1. The total variance explained was 71.961%. Considering that 60% of the variance explained in scales with more than one dimensions is considered sufficient (Hinkin, 1998; Hair et al., 2010), the high total variance ratio we obtained indicates that the factor structure of the test is strong.

When the factor loadings of the TNL items are analyzed, it is seen that the factor loadings vary between .25 and .97 (see Table 4). In order to decide whether an item is related to the conceptual structure or not, it is suggested that the factor loading of that item needs to be at least .30 (Hopkins, 2000; Şencan, 2005; Büyüköztürk, 2020). It is seen that the factor loadings of 95 items in the test are higher than .30. Büyüköztürk (2013) reported that items with factor loadings of .30 and higher discriminate individuals well, items between .20-.30 can be removed from the test if deemed necessary or the item should be corrected, and items with factor loadings lower than .20 should be removed from the test. Anastasia and Urbina (1997), and Child (2006) suggest that items can be tolerated if necessary, unless the factor loading value is below 0.20. In this regard, two items (item 1 and item 76) with factor loadings between .20 and .30 were not removed from the test. Accordingly, no items were removed from the test and the test consisted of 97 items in total. As a result, it can be said that the discrimination of the scale items is quite high.

Confirmatory factor analysis was conducted to determine whether the factor structure in the original form of the scale could be confirmed in a sample of 5-year-old Turkish children. For confirmatory factor analysis, Chi-Square (X2), Degrees of Freedom (df), the ratio of Chi-Square to Degrees of Freedom (X2/df), and RMSEA, GFI, AGFI, CFI and NNFI fit coefficients were calculated.

Şimşek (2007) suggests that the acceptable degree of fit differs for each index. There is no standard interpretation for X2 and df, but in general, smaller values indicate a more accurate fit. In the analysis, the X2 value was 1572.36 and the df was 1021. The value obtained with X2/df is mostly used in determining the fit of the model. When these values are compared to each other (X2/df; 1572.36 /1021), the result is 1.540 (see Table 5). A value of 3 or less indicates that the model has a good goodness of fit, while a value of 5 or less indicates that the model has an acceptable goodness of fit (Çokluk et al., 2021). Therefore, it can be concluded that the obtained value indicates a good fit.

Brown (2006) suggests that an RMSEA value below 0.06 is a good fit for the model, and a value below 0.08 is acceptable. The RMSEA value calculated in the present study is 0.021 and according to this result, the model shows a good fit. For acceptable goodness of fit, GFI, CFI and NNFI values should be .90 or higher and AGFI value should be .80 or higher (Kline, 2010). The analysis showed that GFI was 0.96, CFI was 0.97, NNFI was 0.97, and AGFI was 0.89. The data from confirmatory factor analysis indicated that the goodness of fit of the model was acceptable. In other words, the CFA results demonstrate that the model exhibits a good fit. The fit index values resulting from the CFA indicate that the scale items were appropriately selected for the subtests.

The age factor was also taken into consideration to reveal the construct validity of the TNL. Children's mean scores obtained from the tasks in the test increased with age. The difference between the mean scores of two age groups (60-66 months and 67-72 months) was significant at the .05 level.

As in other language domains, pragmatic (language use) development is recognized to increase with age. Research in the field of pragmatic development has shown that age has effects on children's pragmatic development: children's ability to answer questions and provide complex contextual information for answers, and their level of understanding of grammatical structures and words increase and improve with age (Ryder & Leinon, 2003; Güler & Baykoç Dönmez, 2007). In the studies conducted, it was concluded that the stories told by children develop with age (Karabaş 2002; Çelikli 2020; Khan et al., 2016). Eriksson (2006) states that age has a significant effect on the size of vocabulary and average length of speech. Kanmaz (2019) reported that the number of different words, total number of words, and average sentence lengths increase with age. In another study, Kosaka (2016) compared children aged 4 and 5 years and found that children aged 5 years and above were able to produce stories richer in terms of structure and content and that different storytelling skills were exhibited for each age in the preschool period. In this respect, considering that age is a characteristic determinant of children's receptive and expressive language skills, it can be suggested that the data presented in Table 6 support the construct validity of the TNL. As a result, children are naturally expected to be more successful in receptive and expressive language skills as they get older.

Reliability

Reliability refers to the ability of an instrument to measure consistently (Tavakol & Dennick, 2011). That is, reliability tells us how consistently, invariably or stably a method measures something. When we apply the same method to the same sample under as similar or identical conditions as possible, we are expected to get the same results. Otherwise, the measurement method may be unreliable and the results and scores obtained from a measurement tool with poor reliability will be lacking in credibility (Öner, 1997).

Four types of reliability (internal consistency, split-half test, test-retest, and inter-rater reliability) were used to determine whether the results of the Turkish version of the TNL are reliable and if so, to what extent. Looking at the original form of the test, it is seen that three types of reliability were used: internal consistency, test-retest and inter-rater reliability (Gillam & Pearson, 2004). In general terms, internal consistency refers to the overall agreement between items, and split-half test reliability refers to the correlation between the scores of the two halves of the scale divided into two parts. Test-retest reliability refers to the degree to which a test produces similar results over time, and inter-rater consistency refers to the degree of agreement or consistency between the scores of two or more raters.

Since the TNL has a triple scoring system, Cronbach's Alpha formula was used to calculate internal consistency. Alpha coefficient was developed by Cronbach as a generalized measure of the internal consistency of a multi-item (Likert-type) scale (Peterson, 1994). In other words, it assesses the degree to which items in a test are related to each other.

Alpha ranges between 0 and 1 and a minimum reliability threshold of 0.70 is recommended (Cortina, 1993; Frost et al., 2007; Büyüköztürk, 2013). High alpha values indicate a high degree of correlation between items in a test (Tavakol & Dennick, 2011). However, caution should be taken when interpreting alpha values and it is important to remember that alpha is affected by the number of items in a test because the more items in a test, the higher the alpha value. In fact, values higher than 0.95 do not always indicate high reliability because this may indicate the presence of redundant items in the test (Hulin et al., 2001).

The Cronbach Alpha (α) coefficient calculated for the internal consistency of the TNL in this study was found to be 0.78. Şencan (2005) stated that the reliability value calculated for the overall test may be lower and emphasized that if a test consists of subtests, the alpha coefficient should be calculated separately for each subtest. In the study, the Cronbach Alpha (α) coefficients for the tasks of the test ranged between 0.73 and 0.87. The internal consistency coefficients calculated for the tasks one to six are as follows: .76, .79, .73, .78, .78, and .87. The fact that these values are above the minimum reliability threshold of .70 and below the .95 level, which may indicate the presence of unnecessary items in the test, can be considered as an indicator that the test and the results from the test are reliable.

On the other hand, Şencan (2005) emphasizes that Cronbach's Alpha value may not be strong enough for multidimensional scales and is a good reliability coefficient only for unidimensional scales, adding that it would not be correct to use the alpha value in multidimensional scales on its own to reveal the reliability of the entire scale. Therefore, it would be a more appropriate approach to use different methods to ensure reliability.

One of the ways to estimate the reliability of a measurement tool is to use the same tool to measure the same thing at two different points in time. In psychometrics, this approach is called the test-retest method (Cohen & Swerdlik, 2018). Test-retest reliability refers to the ability of a measurement tool to produce the same results for the same participants when repeated in different situations under the same conditions (Berchtold, 2016) and the high correlation between the scores obtained from two measurements (Baykul, 2021).

In the test-retest analysis, the correlation coefficients between the data obtained from the previous and subsequent measurements are calculated. If the correlation coefficient is high, this is considered evidence of test-retest reliability. In other words, the smaller the difference between the two results, the higher the test-retest reliability. The correlation coefficient is a value ranging from -1.00 to +1.00 and the correlation (consistency/stability) coefficient should be close to +1 for reliability.

However, the test-retest procedure makes the assumption that the measured trait does not change over time. If subjects in a study change at different times between the first and second measurement in terms of the trait being measured, the correlation between the two points in time may be low, even if the measurement instrument is highly sensitive (Collins, 2007). Several factors can influence measurement results at different points in time. For example, subjects may learn new things, forget things, acquire new skills, or external circumstances may affect their ability to respond correctly. The length of time can be a source of error variance. The longer the time elapsed, the more likely the reliability coefficient will be low (Cohen & Swerdlik, 2018). Test-retest reliability can be used to assess how well a method withstands these factors over time.

Based on the literature, in order to calculate the test-retest reliability of the Turkish version of the TNL, a period of 4 weeks between the two test administrations was deemed appropriate. The test-retest reliability conducted with 20 children, corresponding to 12% of the total number of participants, revealed correlation coefficients ranging from .83 to .87 for the six tasks in the test. The test-retest coefficients for tasks one to six were .83, .84, .84, .84, .86, .86, and .87, respectively. In the original form, there was no correlation coefficient calculated specifically for the 5-year age group, and it was calculated with 27 children aged between 5 and 10 years and correlation coefficients of .82 (narration) and .85 (comprehension) were found and these values are close to the correlation coefficients obtained in this study.

Split-half test reliability is another widely used statistical method to measure the internal consistency reliability of a test. It involves dividing a test into two halves and correlating the scores obtained from the two halves, as the name suggests, which requires the test to be administered only once (Thompson, 2010; Frey, 2018). When calculating split-half test reliability, the Spearman-Brown formula is commonly used to estimate full test reliability from the split-test correlation. The Spearman-Brown formula roughly estimates how much the reliability of test scores will change depending on the number of observations or items in a test (Frey, 2018).

When calculating reliability coefficients, there are several ways to split a test. Simply dividing the test into halves is not recommended as it is likely to falsely raise or lower the reliability coefficient (Cohen & Swerdlik, 2018). A test can be divided into two halves by randomly assigning items to one or the other half of the test, by assigning odd-numbered items to one half of the test and even-numbered items to the other half, or by dividing the test by content so that each half contains items that are equivalent in content and difficulty (Frederic, 1956; Crocker & Algina, 1986). The aim here is to create mini-parallel forms in which one half is equal or nearly equal to the other.

Since the TNL contains six tasks, each task was considered as a sub-dimension and each sub-dimension was divided into two halves unbiasedly and the reliability coefficients were calculated. In fact, Tavşancıl (2019) reported that the split-half test reliability is the most widely used method among the methods used to determine scale reliability, and that if the scale has dimensions, these can be considered as a whole within itself and this can also be done for the dimensions. As a result of the calculations, it was seen that the correlation coefficients for the tasks ranged from .74 to .83 (see Table 9). Since the values between 0.70 and 0.89 indicate a strong relationship, it can be concluded that the two-half test reliability of the instrument is high.

Inter-rater reliability was examined as the last step to ensure the reliability of the TNL. Inter-rater reliability is defined as the degree to which two or more raters get the same results under similar assessment conditions (Kottner et al., 2011) and, it can be used to assess the consistency of observations and is useful for data interpretation compared to reliability measures. Inter-rater reliability can be a concern to some extent in many comprehensive studies due to the possibility that multiple data collectors may experience and interpret the target behavior or situation differently. Different observers naturally have different perceptions of situations and events. In reliable research, subjectivity is minimized as much as possible so that a another researcher can reach the same results. When designing the scale and criteria for data collection, different people are expected to consistently assess the same variable with minimal bias. This is particularly important when there is more than one researcher involved in data collection or analysis. Therefore, well-designed research studies should include procedures that measure agreement between raters (McHugh, 2012).

When analyzing the relationship between scorings, it is generally preferred that the calculated correlation or fit coefficient is 0.70 and above, and that it is as close to +1.00 as possible (Erkuş, 2019). When the correlation coefficient is close to +1.00, it is interpreted that different raters score the answers in the test in a similar way and there are few errors in scoring. A weak relationship between the scores or inconsistent scores indicate that the scoring reliability is low (Çetin, 2019).

To ensure the inter-rater reliability of the Turkish version of the TNL, the data from 10 randomly selected children to whom the researcher administered the test were re-scored by an academician who is an expert in the field of child development and education. The scores regarding the dimensions of the scale were compared with the researcher's scores and analyzed using the Spearman Brown formula. The results of the analysis showed that the correlation coefficients ranged from .82 to .90 for the six tasks (see Table 10). The calculated correlation or fit coefficients of 0.70 and higher confirm the inter-rater reliability of the Turkish version of the TNL.

In the present study, cultural adaptation of the TNL for 5-year-old (60-72 months) Turkish children and its validity and reliability were conducted. Considering that the original form of the test was developed to measure the narrative skills of children between the ages of 5-12, in future studies, Turkish adaptation, validity and reliability for other age groups will allow a wider age range to be reached in the assessment of children's language skills through stories. In this sense, it would be beneficial to conduct longitudinal studies on the importance of narrative skills and monitor the effects of various variables on narrative development by using the TNL to follow the narrative skills of children starting from the age of 5 until the age of 12.

The TNL can be used in the assessment of children with and without developmental language disorders. This study included children with normal language development. In the future, in order to distinguish between children with adequate spoken language development and children with developmental language disorders, studies including children with language development problems can be conducted and the instrument can be used to monitor the development of these children and determine the strengths and weaknesses of their spoken language skills.

Family is another important factor in children's language development. In early years, parents are of great importance for children to acquire various language skills and increase their proficiency in language development. The TNL can be used to see the possible effects of variables on children's oral language development such as shared reading, daily reading time, pre- and post-reading activities, and parental attitudes towards reading.

TNL can serve as an example and criterion in the development of new measurement tools that can assess children's oral language skills through stories and in the adaptation of existing measurement tools into Turkish.

It can contribute to the data collection and evaluation process by utilizing it together with other measurement tools intended for this purpose in various studies to be conducted to evaluate the language development of preschoolers.

REFERENCES

Anastasia, A. & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice-Hall

Başol, G. (2019). *Eğitimde ölçme ve değerlendirme* (6. Baskı). Ankara: Pegem Akademi Yayıncılık.

Baykul, Y. (2021). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması*. (4. Baskı) Ankara. Pegem Akademi. DOI 10.14527/9786053640882

Beaton, D., Bombardier, C., Guillemin, F. & Ferraz, M. (2001). Guidelines for the Process of Cross-Cultural Adaption of Self-Report Measures. Spine. 25. 3186-91. 10.1097/00007632-200012150-00014.

Berchtold, A. (2016). Test–retest: Agreement or reliability? Methodological Innovations. https://doi.org/10.1177/2059799116672875

Boudreau, D. (2008). Narrative abilities - Advances in research and implications for clinical practice. *Topics in Language Disorders*, 28(2), 99-114.

Bower, G. (1976). Experiments on story understanding and recall. *Quarterly Journal of Experimental Psychology*, 28,5 1 1-534

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research series.* New York, USA: The Guilford Press.

Büyüköztürk Ş. (2020). *Sosyal Bilimler İçin Veri Analizi El Kitabı. İstatistik, Araştırma Deseni SPSS Uygulamaları ve Yorum* (28. Baskı) Ankara: Pegem Akademi, DOI 10.14527/9789756802748

Büyüköztürk, Ş. (2013). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem Yayıncılık.

Child, D. (2006). The Essentials of Factor Analysis. 3rd edn. New York: Continuum.

Cohen, R. J. & Swerdlik, M. E. (2018). *Psychological testing and assessment : an introduction to tests and measurement* (Ninth edition.). McGraw-Hill Education.

Collins, L.M. (2007). *Research Design and Methods*. Encyclopedia of Gerontology (Second Edition), Pages 433-442,

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104. doi:10.1037/0021-9010.78.1.98.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt. Rinehart and Winston

Çelikli, C.B. (2020). 6-8 yaş arası çocukların dil becerilerinin öyküleme yoluyla değerlendirilmesi. Yüksek LİsansTezi, Ege Üniversitesi Sağlık Bilimleri Enstitüsü, İzmir.

Çetin, B. (2019). *Eğitimde Ölçme ve Değerlendirme* (1. Baskı). Ankara: Arı Yayıncılık

Çilingirtürk A.M. (2011). İstatistiksel Karar Almada Veri Analizi. (1. Baskı) Ankara: Seçkin Yayıncılık

Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2021). Sosyal Bilimler İçin Çok Değişkenli İstatistik: SPSS ve Lisrel Uygulamaları. (6. Baskı) Ankara: Pegem Akademi Yayıncılık.

Dale, P. & Henderson, V. (1987). An Evaluation of the Test of Early Language Development as a Measure of Receptive and Expressive Language. Language Speech and Hearing Services in Schools. 18. 179. 10.1044/0161-1461.1802.179.

Davies, A. (2007). Storytelling in the classroom, enhancing traditional oral skills for teachers and pupils. SAGE Publications (Paul Chapman Publishing): London, UK

Demir, S. (2017). Ölçmede Geçerlik. (Editör: R. N. Demirtaşlı). Eğitimde Ölçme ve Değerlendirme (4. Baskı). Ankara: Anı Yayıncılık (s. 57-76)

Demircioğlu, G. (2015). Geçerlilik ve Güvenilirlik (Editör: Emin Karip) Ölçme ve Değerlendirme (7. Baskı) Ankara: Pegem Akademi Yayıncılık (s. 89-122)

Duinmeijer, I., de Jong, J. & Scheper, A. (2012). Narrative abilities, memory, and attention in children with a specific language impairment. *International Journal of Language and Communication Disorders, 47*(5), 542-55

Eriksson, M. (2006). Sex differences in language development as a topic for cross-culturalcomparisons. *Proceedings from the First European Network Meeting on the Communica-tive Development Inventories*,103–114.

Erkuş, A. (2003). *Psikometri üzerine yazılar.* Ankara: Türk Psikologlar Derneği Yayınları. No 24.

Erkuş, A. (2019). *Psikolojide Ölçme ve Ölçek Geliştirme I: Temel Kavramlar ve İşlemler* (4. Baskı). Ankara: Pegem Yayınları

Frederic, M. L. (1956). Sampling Error due to Choice of Split in Split-Half Reliability Coefficients, The Journal of Experimental Education, 24:3, 245-249, DOI: 10.1080/00220973.1956.11010545

Frey, B. (2018). The SAGE encyclopedia of educational research, measurement, and evaluation (Vols. 1-4). Thousand Oaks,, CA: SAGE Publications, Inc. doi: 10.4135/9781506326139 Retrieved 15.06.2021 from https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i19572.xml.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures?. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research, 10 Suppl 2, S94–S105. https://doi.org/10.1111/j.1524-4733.2007.00272.x

Gamble, N. & Yates, S. (2008), Exploring Children’s Literature, London: Sage

Gambrell, L. B. & Dromsky, A. (2000). Fostering reading comprehension. In D. S. Strickland & L. M. Morrow (Eds.), *Beginning reading and writing* (pp. 143-153). New York, NY: Teachers College Press.

Gillam, R. B. & Pearson, N. (2004). Test of Narrative Language. Austin, TX: PRO-ED

Golden, J. (1984). Children's concept of story in reading and writing. *Reading Teacher, 37,* 578-584.

Güler, T. & Baykoç Dönmez, N. (2007). "*48-72 Aylar Arasındaki Türk Çocuklarının Alıcı Dil Yapılarının İncelenmesi*", *Eğitim Araştırmaları Dergisi*, 7 (27), 83-96.

Hair, J., Black, W., Babin, B. & Anderson, R. (2010). Multivariate data analysis (7th ed.). Upper Saddle River, Prentice-Hall, NJ, USA.

Hambleton, R. K. (2017). Testlerin Birden Çok Dil ve Kültüre Uyarlanmasıyla İlgili Konular, Uyaralama Desenleri ve Uyarlama İçin Teknik Yönergeler (Ed. Ronald K. Hableton, Peter F. Merenda, Charles D. Spielberg), İçinde, *Eğitimde Ve Psikolojide Kullanılan Testlerin Kültürlerarası Değerlendirme Amacıyla Uyarlanması* (s. 1-40).

Heilmann, J., Miller, J. F., & Nockerts, A. (2010). Sensitivity of narrative organization measures using narrative retells produced by young school age children. *Language Testing*, 27(4), 603-626

Hinkin, T. R. (1998). A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires. Organizational Research Methods, 1(1), 104–121. https://doi.org/10.1177/109442819800100106

Hopkins, W. G. (2000). A scale of magnitudes for effect statistics. Retrieved 28.06.2021 from http://www.sportsci.org /resource/stats/effectmag.html

Hudson, J.A., & Shapiro, L.R. (1991). From knowing to retelling: The development of children's scripts, stories, and personal narratives. In McCabe, A. & Peterson, C. (Eds.), Developing Narrative Structure. Lawrence Erlbaum Associates, Inc. Hillsdale

Hulin, C. & Netemeyer, R. & Cudeck, R. (2001). Can a Reliability Coefficient Be Too High? *Journal of Consumer Psychology*. 10. 55-58. 10.2307/1480474.

Justice, L. M., Bowles, R., Kaderavek, J., Ukrainetz, T., Eisenberg, S.,& Gillam, R. (2006). The Index of Narrative Microstructure: A clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*,15, 177-91. 10.1044/1058-0360(2006/017).

Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39(1), 31–36. https://doi.org/10.1007/BF02291575

Kalfoss, M. (2019). Translation and Adaption of Questionnaires: A Nursing Challenge. SAGE Open Nursing. 5. 237796081881681. 10.1177/2377960818816810.

Kan, A. (2019). Ölçme Araçlarında Bulunması Gereken Nitelikler. (Editör: H. Atılgan). Eğitimde Ölçme ve Değerlendirme (12. Baskı) Ankara: Anı Yayıncılık (s. 43-102)

Kanmaz, S. (2019). Okul öncesi dönemdeki çocukların öyküleme yoluyla dil becerilerinin değerlendirilmesi. Yüksek LİSans Tezi, Ege Üniversitesi Sağlık Bilimleri Enstitüsü, İzmir.

Kapuscinski, A. N. & Masters, K. S. (2010). The current status of measures of spirituality: a critical review of scale development. Psychology of Religion and Spirituality, 2(4), 191–205. http://dx.doi.org/10.1037/a0020498.

Karabaş, Ö. (2002). Okul öncesi dönemindeki Türk çocuklarının hikâye öyküleme becerisinin gelişimi. Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi / Sosyal Bilimler Enstitüsü, Ankara.

Karasar, N. (2009). *Bilimsel araştırma yöntemi: kavramlar, ilkeler, teknikler*. Nobel Yayın

Karasar, N. (2013). *Bilimsel araştırma yöntemi.* Ankara: Nobel Yayın Dağıtım.

Khan, K., Gugiu, M., Justice, L., Bowles, R., Skibbe, L. & Piasta, S. (2016). Age-Related Progressions in Story Structure in Young Children's Narratives. *Journal of Speech Language and Hearing Research. 59.* 1. 10.1044/2016_JSLHR-L-15-0275.

Kline, P. (2000) *The handbook of psychologial testing.* London: Taylor & Francis Group

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kosaka, M. (2016). Narrative Development of Contents and Structures in Story Telling. *The Japan Journal of Logopedics and Phoniatrics. 57. 261-271. 10.5112/jjlp.57.261.*

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Journal of clinical epidemiology, 64(1), 96–106. Retrieved 16.06.2021 from https://doi.org/10.1016/j.jclinepi.2010.03.002

Leahy, M. J., Szymanski, E. M. & Linkowski, D. C. (1993). Knowledge importance in rehabilitation counseling. *Rehabilitation Counselling Bulletin, 37,* 130-145.

Mages, W. K. (2008). Language and theory of mind development in the context of a Head Start Theatre-in-Education program. PhD dissertation., Harvard University.

Mandler, J. M. & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, vol. 9 (January 1977), pp. 111-151.

Martin, R. A. & Ford T.E. (2018). The Personality Psychology of Humor, (Editors: Rod A. Martin, Thomas E. Ford). The Psychology of Humor (p. 99-140) Academic Press, (Second Edition) ISBN 9780128121436, https://doi.org/10.1016/B978-0-12-812143-6.00004-7.(https://www.sciencedirect.com/science/aticle/pii/B9780128121436000047)

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

McIntyre, L. J. (2005). Investigating teachers ‘knowledge of oral language, Doctoral Dissertation. University of Alberta, Canada.

MEB (2013). *Milli Eğitim Bakanlığı okul öncesi eğitim programı. http://docplayer.biz.tr/895440-T-c-milli-egitim-bakanligi-talim-ve-terbiye-kurulu-baskanligi-konu-okul-oncesi-egitim-programı.html.*

Meier, A. (2020). Early growth trajectories of narrative language development: Sex differences in microstructural and macrostructural skills (Order No. 27958836). Available from ProQuest Dissertations & Theses Global. (2399400084). Retrieved 25.04.2021 from *https://search.proquest.com/dissertations-theses/early-growth-trajectories-narrative-language/docview/2399400084/se-2?accountid=16935*

Merenda, P. F. (2017). Eğitimde ve Psikolojide Kullanılan Testlerin Kültürlerarası Uyarlanması (Ed. Ronald K. Hableton, Peter F. Merenda, Charles D. Spielberg), İçinde, *Eğitimde Ve Psikolojide Kullanılan Testlerin Kültürlerarası Değerlendirme Amacıyla Uyarlanması* (s. 329-350). Nizamettin Koç ve Ahmet Yıldırım (Çev.) Ankara: Pegem Akademi Yayıncılık

Merritt, D.D. & Liles B.Z. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling, and story comprehension. *Journal of Speech and Hearing Research, 30,* 539–552.

Milburn, T.F., Girolametto, L., Weitzmam, E. & Greenberg, J. (2014) Enhancing preschool educators’ ability to facilitate conversations during shared book reading. Journal of Early Childhood Literacy 14: 105–40.

Mills, M. T. (2015). The Effects of Visual Stimuli on the Spoken Narrative Performance of School-Age African American Children. Language, speech, and hearing services in schools, 46(4), 337–351. https://doi.org/10.1044/2015_LSHSS-14-0070

Morrow, L. (1985). Retelling stories: A strategy for improving young children’s comprehension, concept of story structure, and oral language complexity. *The Elementary School Journal, 85*, 647-661.

Morrow, L. M. (1986}. Effects of structural guidance in story retelling on children's dictation of original stories. Journal of Reading Behavior, 18, (2), 135-152.

Morrow, L. M. (1990). Assessing children’s understanding of story through their construction and reconstruction of narrative. In L. M. Morrow & J. K. Smith (Eds.), Assessment for instruction in early literacy (pp. 110–134). Englewood Cliffs, NJ: Prentice-Hall

Öner, N. (1997). *Türkiye’de kullanılan psikolojik testler: Bir başvuru kaynağı* (3.baskı). İstanbul: Boğaziçi Üniversitesi Yayınları

Özgüven, İ.E. (2011). *Psikolojik testler*. Ankara: Pdrem Yayınları.

Petersen, D. B., Gillam, S. L., Spencer, T., & Gillam, R. B. (2010). The effects of literate narrative intervention on children with neurologically based language impairments: An early stage study. *Journal of Speech, Language, and Hearing Research, 53*, 961-981.

Peterson, C. (1990). The who, when, and where of early narratives. *Journal of Child Language*, *17*, 433-455.

Peterson, R. (1994). A Meta-Analysis of Cronbach's Coefficient Alpha. Journal of Consumer Research. 21. 381-91. 10.1086/209405.

Richert, R., Sachet, A., Hoffman, R. & Taylor, M. (2009). Learning From Fantasy and Real Characters in Preschool and Kindergarten. *Journal of Cognition and Development*. 10. 41-66. 10.1080/15248370902966594.

Ryder, N. & Leinonen, E. (2003). Use of Context in Question Answering by 3-, 4- and 5-Year-Old Children. *Journal of psycholinguistic research. 32*. 397-415. 10.1023/A:1024847529077.

Saad, S., Carter, G., Rothenberg, M., & Israelson, E. (1999). Testing and Assessment: An Employer's Guide to Good Practices.

Seçer, İ. (2015). Psikolojik Test Geliştirme ve Uygulama Süreci. (1. Baskı). Anı Yayıncılık: Akara.

Şencan, H. (2005). *Sosyal ve davranissal ölçümlerde güvenilirlik ve geçerlilik.* (1. Baskı). Ankara: Seçkin Yayınevi.

Snaith, M. (2007). Children's Care Learning and Development, London: Heinemann

Stadler, M. & Ward, G. (2005). Supporting the Narrative Development of Young Children. *Early Childhood Education Journal. 33*. 73-80. 10.1007/s10643-005-0024-4.

Şimşek, Ö. F. (2007). *Yapısal eşitlik modellemesine giriş: Temel ilkeler ve LISREL uygulamaları.* Ankara: Ekinoks.

Tatlıdil, H. (2002). *Uygulamalı çok değişkenli istatistiksel analiz*. Ankara: Akademi Matbaası

Tavakol, M. & Dennick, R. (2011). Making Sense of Cronbach's Alpha. International Journal of Medical Education. 2. 53-55. 10.5116/ijme.4dfb.8dfd.

Tavşancıl, E. (2019). *Tutumların Ölçülmesi ve SPSS ile Veri Analizi* (6.Baskı) Ankara: Nobel Akademik Yayıncılık.

Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the Maximal Split-Half Coefficient to Estimate Reliability. Educational and Psychological Measurement, 70(2), 232–251. https://doi.org/10.1177/0013164409355688

Tyupa, S. (2011). A theoretical framework for back-translation as a quality assessment tool. 7. 35-46.

Uzunboylu, H., & Ozdamlı, F. (2011). Teacher perception for m-learning: scale development and teachers' perceptions. Journal of Computer Assisted Learning, 27, 544–556. http://dx.doi.org/10.1111/j.1365-2729.2011.00415.x.

van Kraayenoord, C. E., & Paris, S. G. (1996). Story construction from a picture book: An assessment activity for young learners. Early Childhood Research Quarterly, 11(1), 41–61

Walker, C. M., Ganea, P. A. & Gopnik, A. (2012). Children's Causal Learning from Fiction: Assessing the Proximity Between Real and Fictional Worlds. Proceedings of the Annual Meeting of the Cognitive Science Society, 34. Retrieved 26.04.2021 from https://escholarship.org/uc/item/3dk4b2tr

Weisberg, D., Ilgaz, H., Hirsh-Pasek, K., Golinkoff, R., Nicolopoulou, A. & Dickinson, D. (2015). Shovels and swords: How realistic and fantastical themes affect children's word learning. Cognitive Development. 35. 10.1016/j.cogdev.2014.11.001.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A. & Erikson, P. (2005), Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in Health, 8: 94-104. https://doi.org/10.1111/j.1524-4733.2005.04054.x