IJMS
Vol 45, No 6, November 2020

Original Article

The Development of a New Test for Consecutive Assessment of Narrative Skills in Iranian School-Age Children

Saeide Beytollahi[1], MSc; Zahra Soleymani[1], PhD; Shohreh Jalaie[2], PhD

[1]Department of Speech Therapy, School of Rehabilitation, Tehran University of Medical Sciences, Tehran, Iran;
[2]Department of Physiotherapy, School of rehabilitation, Tehran University of Medical Sciences, Tehran, Iran

Correspondence:
Zahra Soleymani, PhD;
Department of Speech Therapy, School of Rehabilitation, Tehran University of Medical Sciences, Enghelab Ave., Pitch-e-Shemiran, Postal code: 11489-65111, Tehran, Iran
Tel: +98 9122036683
Fax: +98 21 77534133
Email: soleymaniz@sina.tums.ac.ir
Received: 09 April 2019
Revised: 10 June 2019
Accepted: 23 June 2019

Abstract

Background: The assessment of narrative skills in young children is essential for early identification of potential learning and literacy-related difficulties. The present study aimed to develop a validated and reliable test for consecutive assessment of narrative skills in Iranian school-age children.

Methods: To elicit story retelling by the children, 14 pictures (each a discrete test) were developed based on which 14 stories were scripted in accordance with the Stein and Glenn model. The pictures were presented to 50 typically developing (TD) children and seven children with autism spectrum disorder (ASD) in Kerman, Iran, 2018. The recorded audio of retold stories by the children was transcribed and analyzed using the monitoring indicator of scholarly language (MISL) instrument. The SPSS software (Version: 18.0) was used to analyze the data with the significance level set at 0.05.

Results: A high correlation between the total MISL score of each test and each MISL item ($r>0.5$) confirmed the construct validity of our test. A comparison of the mean total MISL score between the TD and ASD groups showed significant differences ($P<0.001$) for all pictures. The internal consistency coefficient was >0.7 between all the MISL items and the intraclass correlation coefficient between the test and retest scores was >0.96 for all pictures. The repeated measures ANOVA did not indicate significant differences ($P=0.15$) between the total MISL scores of the pictures.

Conclusion: The validity and reliability of the developed test were confirmed, suggesting it can be used for consecutive assessment and monitoring of the narrative skills in school-aged children.

Please cite this article as: Beytollahi S, Soleymani Z, Jalaie S. The Development of a New Test for Consecutive Assessment of Narrative Skills in Iranian School-Age Children. Iran J Med Sci. 2020;45(6):425-433. doi: 10.30476/ijms.2019.81984.

Keywords ● Language tests ● Narration ● Autism spectrum disorder ● Psychometrics

What's Known

- Assessment of narrative discourse in children with language impairments is essential.
- Several instruments are available to assess different types of narratives, including retelling and self-generated stories. However, none of these tools are suitable for the consecutive assessment of narrative skills.

What's New

- For the first time, a new test is developed for consecutive assessment and monitoring of narrative skills in school-age children.
- The validity and reliability of the test were confirmed. Discriminant validity of the test was confirmed following testing with typically developing and autistic children.

Introduction

Most parents and teachers inspire preschool children to talk about events that occurred during the day or in a distant past.[1] Narration is an important aspect of language and defined as stories about real or imaginary events with a causal relationship or which are in temporal order.[2] Narrative discourse plays an important role in children's socio-cultural and language development and is considered a predictor of literacy achievement, particularly

reading comprehension.[3, 4] It is therefore essential that children understand and effectively produce appropriate narrative forms.[5] There are three types of narrative assessment methods, namely story comprehension, story retelling, and personal narrative. Story comprehension involves reading a story to a child and asking questions about the story. In story retelling, an examiner reads a model story to a child, and the child should retell the story. Finally, personal narrative requires children to formulate stories from pictures or telling their personal story that is thematically related to a model story. Retelling and personal narrative provide different insights about children's storytelling skills.[6]

Story grammar model is a common way for macro-structural analysis of narrative. According to this model, each story has a setting and an episode system.[7] The setting provides information about the time or the place where the story had occurred. The episode system comprises three elements. The first element is a problem or an initiating event and/or internal response, which is an occurrence that motivates actions by the main character(s). The second, goal-directed actions for solving the problem known as attempts and the third element is the consequence (or outcome) associated with both the initiating event and the actions.[8] Another area for the development of narratives is literate language and cohesive devices. Literate language is an essential area of narrative microstructure.[9] Narrative competency is using elaborated noun phrases, metalinguistic verbs such as say or tell, and metacognitive verbs namely think or believe. They are main literate language components.[10] Narrative skills are diminished in children with specific language impairments and those with language/learning disabilities.[11] In order to discriminate between children with typical development (TD) and children with language disorders, macrostructural analysis has been used.[12]

There are several instruments for assessing narrative performance. These instruments are divided into two groups, namely norm-referenced and criterion-referenced. Norm-referenced tests are designed to compare one child with another, based on which a therapist can rapidly and precisely compare children's narratives to the peers. Criterion-referenced tests are used to measure the knowledge and skills of a child against a set of predetermined criteria.[13] Two standardized norm-referenced instruments including the Renfrew bus story and the test of narrative language are available for assessment of narratives.[2] The narrative "Frog, where are you?"[14] and the Persian narrative norms instrument[15] are the commonly known criterion-referenced methods in Iran for analyzing language content, form, and use in children. In these methods, a child is asked to tell a story based on a wordless picture book from which a language sample is analyzed.[16] Checklists, rating scales, or rubrics are mainly applied as tools to analyze the results of criterion-referenced tests, such as monitoring indicators of scholarly language (MISL)[17] and narrative scoring scheme (NSS).[5]

The present study aimed to develop a valid and reliable criterion-referenced test. Compared to other narrative assessment tests, this newly developed test is specifically designed for consecutive assessment of spontaneous narratives in young school-age children in a short period of time. It is known that when clinicians or researchers use the same test repeatedly for progress monitoring, the results may not be reliable due to familiarization. The uniqueness of our test is that it consists of several independent tests for progress-monitoring, all of which assess the same item and the corresponding scores are comparable with one another. Furthermore, criterion-referenced tests facilitate intervention planning, since they indicate the weaknesses and strengths of a specific area in detail.

Materials and Methods

Participants

Two groups of children were recruited in the study, including typically developing children (TD) as the control group (n=50) and children with autism spectrum disorder (ASD) as the group with language impairments (n=7). The TD children were recruited from public schools in Kerman, Iran, 2018. Based on a general guideline, a sample size of at least 50 participants was considered adequate for the assessment of the psychometric parameters.[18] The inclusion criteria for the TD children were average achievement scores in core subjects at school and no history of hearing, visual, or gross neurological impairment, as well as psychiatric diagnosis, or learning disability. Moreover, preschool screening tests also confirmed their normal intelligence level. The parents of these children were approached to inquire about their native language and the history of language development. Subsequently, the inclusion criteria were extended to include native Persian-speaking children and normal language development. The speech-language pathology (SLP) test was performed to confirm normal oral structure with no signs of stuttering, speech errors, language impairment, or attention deficit

disorder in each child in the TD group. As part of the selection process, the teachers at those public schools were informed about the study protocol and requested to identify qualified children.

The ASD children were recruited from private clinics in Kerman, Iran, 2018. Their disorders were confirmed by a qualified child psychiatrist using the DSM-V criteria[19] and the Persian Gilliam autism rating scale, second edition (GARS-2).[20] In the selection process, only boys were considered, since the prevalence of ASD is much higher in boys than in girls,[21] and we could only approach boys. The children in this group were monolingual (Persian) speakers with a nonverbal IQ of ≥85 (89.39±3.15) on the Wechsler preschool and primary scale of intelligence (WPPSI).[22] Due to the high prevalence of language impairments in children with ASD, we used the free-play method to collect language samples from each child. The inclusion criteria were the ability to produce 100 utterances and a mean length of utterances (MLU) of four. The choice for the 100 utterances was to ensure that a low score on the test was not due to their pragmatic disorder. Note that MLU is a useful criterion for indicating syntactic complexity.[23] As a result of the above, the age range of the children in the ASD group was wider that was the TD group.

Besides verbal consent from each child, written informed consent was obtained from the children's parents. The study protocol was approved by the Ethics Committee of Tehran University of Medical Sciences, Tehran, Iran (REC.FNM.TUMS.IR.137.1397).

Pictures and Stories

To elicit story retelling by the children, 14 single-scene pictures (figure 1) as the main test and a single image (figure 2) as an exercise test were developed. The pictures depicted human characters based on which 15 stories were scripted in accordance with the Stein and Glenn model.[7] Each of the 15 stories constituted a discrete test. Since familiarization is crucial in populations such as children, an exercise session was included in the study. All stories were of a single-episode with two or three characters. Each story contained a clear initiating event (e.g., a child getting injured, tearing off clothes, painting on a wall) to set a plot in motion as well as the associated goal-directed actions such as wound dressing and mended torn clothes.

A panel of narrative experts was asked to comment on the developed stories in terms of structure and relevance to children. Based on their feedback, the stories were modified and

Figure 1: A sample of the 14 single-scene pictures that was illustrated.

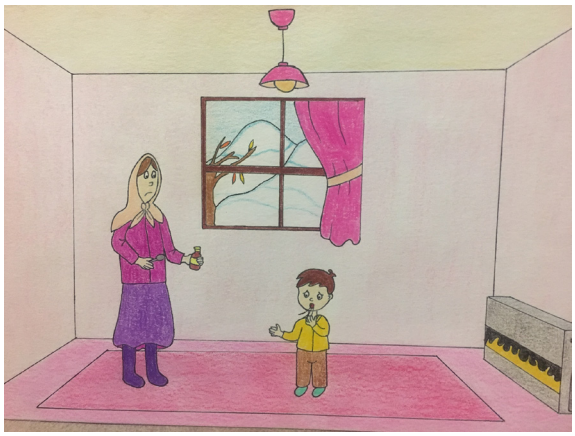

Figure 2: The figure shows one of the pictures used in the exercise session.

subsequently, a professional cartoonist drew the final version of the pictures in accordance with the scripts. The pictures depicted characters, setting (park, house, or doctor's office), initial event, and a problem-solving clue. The face of the human characters was designed to depict an external event in emotions. Except for the main components of the stories, pictures did not show any confusion or extra stimulants. All pictures were colorful and of the same size (A5). The narrative experts reviewed the final version of the pictures and confirmed their appropriateness and representativeness.

Data Collection

Data were collected by three student research assistants in speech-language pathology. The first author fully informed the children as well as their parents about the goal of the study, test method, and the data collection procedure. The children were assessed individually in a quiet room in their own school. The tests were performed in two sessions on two consecutive days, each covering seven pictures (i.e., tests) and lasting approximately 30 minutes. The pictures were randomly presented to the children

to reduce the negative impact of fatigue. At first, an exercise session was carried out using a dedicated picture and its associated story. The picture depicted two main characters (a little boy and his mother) in a room with a window through which a snowy landscape could be seen. The boy seemed to have a cold and coughed, and his mother looked sad and offered him cough medicine. The research assistant presented the picture to each participating child and only gave an overall description (the presence of two individuals and an event). Each child was requested to carefully look at the picture while listening to the pre-defined story (based on the Stein and Glen model). Children were asked to retell the story afterward without seeing the picture to ensure that they would narrate the story as literally as possible and in an understandable fashion. Every effort was made to avoid any association between the exercise session and the main sessions (the 14 single-scene pictures). The main sessions were similarly conducted as the exercise session. To examine test-retest reliability, the 2-day sessions were repeated after one month. With prior permission from the children's parents, recordings of the stories retold by the children were made using the Recordium IOS application software for iPhone. To ensure the correct conduct of the tests, the first author randomly reviewed 20% of all audio recordings using a pre-defined verification checklist.

The recorded data were transcribed verbatim by the research assistants. The transcribed utterances were divided into communication units. Each communication unit included one main clause with all dependent phrase or clauses attached to it.[17] All irrelevant and unintelligible utterances were noted but excluded from the analysis. To ensure correct transcription of the recorded data, a PhD student in speech-language pathology (blinded to the goal of the study) transcribed 10% of the recorded data. The agreement between transcriptions was 98% for C-unit segmentation and 97% for word-by-word agreement (agreement divided by agreement plus disagreement).

The MISL, with its macrostructure and microstructure subscales, was used to analyze the transcribed stories. The macrostructure subscale includes seven story elements (character, setting, initiating event, internal response, plan, action, and consequence). The microstructure subscale contains six items (coordinating conjunction, subordinating conjunction, adverbs, metacognitive verbs, metalinguistic verbs, and elaborated noun phrases). The sum of macrostructure and microstructure scores (total MISL score) represents an index of overall narrative complexity (NC).[17] For each macrostructure subscale, a story element was rated zero when the story did not contain that element. A score of 1 demonstrated that the story included an element that was not directly related to the story. A score of 2 indicated that an element was correctly related to the story, and a score of 3 showed that there were two or more of those elements.[17] The scoring system for the microstructure sub-scale was according to the rating of the items. A score of 0 was given if there were not any items in the story, a score of 1 was given when there was one example of the item in the story, a score of 2 was given when there were two various examples, and a score of 3 was given if three or more various examples were present. To facilitate the scoring, a table was prepared by the first author with the support of a Farsi linguist for the definition of story elements and examples of each element.[17]

Inter-rater Reliability

A PhD student in speech-language pathology (the coder) blind to the goal of the study was trained in the application of the MISL. During a preliminary training, five randomly selected stories were scored by the coder and the results were cross-checked with those of the first author. The scoring method was discussed and potential ambiguities were resolved. Then, the coder and the first author independently scored five additional stories (20% of the total transcripts) and the extent of agreement between the two sets of results was calculated. Since a simple inter-rater agreement does not include the effect of chance, Cohen's kappa coefficient was used to calculate agreement on the NC of each story. The coefficient indicates the agreement between individual scoring categories and the frequency of different categories among raters other than those expected by chance.[24]

Statistical Analysis

Statistical analysis was performed using SPSS software version 18.0. Spearman's rank-order correlation coefficient was used to measure construct validity. Validity was also determined by displaying differences between distinct groups that were predicted empirically.[25] Previous studies have shown that children with ASD have problems with comprehending and producing narratives.[12-26] Therefore, we specifically recruited these children in the present study. Due to the small sample size in the ASD group, the nonparametric Mann-Whitney U test was used to compare the mean NC between the TD and ASD groups. The internal consistency and test-retest reliability were assessed using Cronbach's

alpha and intraclass correlation coefficient (ICC), respectively. Repeated measures ANOVA was used to compare the tests.

Results

The TD group included 50 children (25 boys, 25 girls) with a mean age of 7.7±0.66 years (range 6.6-8.4 years). The ASD group included seven boys with a mean age of 8.8±0.53 years (range: 8.2-9.5 years). The bivariate correlation between the total MISL score of each test and each MISL item confirmed the construct validity of the test. The corresponding Spearman's correlation coefficients are presented in table 1. The results showed a high correlation between the total MISL score of each test and each MISL item (r>0.5).[27] Moreover, the correlation coefficient between the macrostructure and microstructure subscales was between r=69 and r=86 for all stories. In terms of discriminant validity, we anticipated a difference in the total MISL score between the ASD and TD groups. Based on the Mann-Whitney U test, there was a significant difference in the mean NC score of all stories between the two groups (3.95<Z<4, P<0.001). The inter-rater reliability for each story was examined using the kappa coefficient (table 2). The coefficients were significant at P<0.001, indicating an almost perfect inter-rater agreement.[24] We examined reliability via internal consistency reliability. Cronbach's alpha coefficients of the NC scores and the two MISL subscales are presented in table 3. In accordance with a previous study, a reliability coefficient of ≥0.70 was considered acceptable.[27] The internal consistency reliability of the macrostructure subscale was greater than 0.70 in 86% of the pictures, but it was slightly lower for the microstructure subscale of most of the pictures (a>0.60). The ICC (P<0.05) was used to measure test-retest reliability.[28] The ICC values for each story are presented in table 4. According to a subjective guideline provided by Landis and Koch,[29] an ICC value between 0.61 and 0.80 indicates substantial agreement and values between 0.81 and 1.00 indicates almost perfect agreement. Finally, we examined whether all pictures in our instrument assessed the same

Table 1: The correlation coefficient between the total monitoring indicator of scholarly language score of each test and each monitoring indicator of scholarly language item (P<0.001)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
Macrostructures														
Character	0.69	0.86	0.67	0.63	0.73	0.62	0.85	0.85	0.71	0.81	0.72	0.74	0.60	0.79
Setting	0.64	0.69	0.60	0.58	0.65	0.54	0.68	0.64	0.58	0.63	0.60	0.63	0.60	0.58
Initial event	0.73	0.70	0.75	0.72	0.78	0.70	0.72	0.63	0.76	0.69	0.70	0.67	0.61	0.63
Internal response	0.51	0.50	0.52	0.51	0.56	0.52	0.63	0.63	0.51	0.50	0.60	0.56	0.50	0.55
Plan	0.59	0.58	0.51	0.50	0.52	0.50	0.55	0.55	0.53	0.62	0.58	0.57	0.50	0.64
Attempt	0.60	0.52	0.50	0.51	0.56	0.63	0.49	0.50	0.57	0.48	0.58	0.52	0.51	0.52
Consequence	0.63	0.51	0.54	0.67	0.55	0.63	0.69	0.56	0.53	0.58	0.58	0.61	0.56	0.65
Microstructures														
CC	0.69	0.59	0.63	0.52	0.51	0.63	0.62	0.76	0.57	0.55	0.50	0.55	0.51	0.61
SC	0.49	0.57	0.52	0.58	0.51	0.50	0.56	0.48	0.53	0.51	0.52	0.52	0.56	0.50
Mental verbs	0.51	0.49	0.51	0.53	0.52	0.55	0.57	0.60	0.63	0.60	0.63	0.63	0.53	0.63
Linguistic verbs	0.55	0.60	0.59	0.54	0.60	0.49	0.53	0.60	0.51	0.61	0.56	0.51	0.57	0.54
Adverbs	0.64	0.61	0.54	0.69	0.54	0.52	0.49	0.53	0.55	0.52	0.51	0.49	0.51	0.60
ENPh	0.56	0.48	0.51	0.54	0.54	0.63	0.58	0.50	0.57	0.57	0.52	0.54	0.55	0.49

T: Test, CC: Coordinating conjunction, SC: Subordinating conjunction, ENPh: Elaborated noun phrase

Table 2: The Inter-rater agreement for each test between the two raters

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
Cohen's kappa	0.94	0.96	0.93	0.93	0.91	0.95	0.95	0.93	0.92	0.93	0.92	0.94	0.91	0.95

T: Test

Table 3: The internal consistency between the monitoring indicator of scholarly language subscales

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
NC	0.84	0.80	0.83	0.82	0.84	0.83	0.86	0.86	0.85	0.83	0.82	0.83	0.80	0.88
Macrostructures	0.82	0.70	0.75	0.70	0.78	0.73	0.81	0.77	0.75	0.74	0.77	0.75	0.70	0.85
Microstructures	0.66	0.73	0.69	0.63	0.63	0.61	0.62	0.73	0.72	0.63	0.60	0.60	0.68	0.71

NC: Narrative complexity, T: Test

Table 4: The intraclass correlation coefficients between tests and retests

	Test Mean±SD	Retest Mean±SD	ICC	CI	P value
Test 1	21.27±5.10	22.70±4.32	0.97	0.94-0.98	<0.001
Test 2	21.45±5.56	22.90±4.29	0.96	0.94-0.98	<0.001
Test 3	21.74±5.00	22.90±4.24	0.97	0.95-0.98	<0.001
Test 4	21.54±5.12	23.10±4.57	0.97	0.95-0.98	<0.001
Test 5	21.80±5.40	23.02±4.18	0.97	0.96-0.98	<0.001
Test 6	21.69±5.01	22.64±3.86	0.98	0.97-0.99	<0.001
Test 7	21.81±5.21	22.54±4.00	0.97	0.96-0.98	<0.001
Test 8	21.71±5.38	22.62±4.25	0.96	0.94-0.98	<0.001
Test 9	21.67±5.18	22.88±3.94	0.96	0.94-0.98	<0.001
Test 10	21.46±5.43	23.12±4.68	0.97	0.95-0.98	<0.001
Test 11	21.42±5.21	22.70±4.29	0.98	0.96-0.98	<0.001
Test 12	21.39±5.02	22.58±4.20	0.97	0.95-0.98	<0.001
Test 13	21.64±4.75	23.04±4.45	0.97	0.95-0.98	<0.001
Test 14	21.60±5.25	22.78±4.32aq	0.97	0.95-0.98	<0.001

ICC: Intraclass correlation coefficient, CI: Confidence interval (95%). Spearman's correlation coefficients tests were used

item, and if their corresponding scores were comparable. The results of the repeated measures ANOVA (within-subject) showed no significant difference between the 14 tests (F=1.87, P=0.15), and the observed power was 0.741.

Discussion

We developed a valid and reliable test for assessing spontaneous fictional narratives in children aged 6 to 8 years. The test consisted of 14 pictures where each picture considered a separate test. All tests assessed the same item and were administered in the shortest possible time, on average 8-10 minutes. Time is of the essence in narrative assessment, as an assessment test must be performed in a short period of time to adhere to the practice of SLP testing and to reduce the negative impact of fatigue on children.

The recorded audio of the stories retold by the children was transcribed and analyzed using the MISL rubric as the scoring criterion. MISL is used for assessing self-generated narratives extracted from sequenced or single-scene pictures. This scoring system provides the examiners with the opportunity to determine those specific aspects of the narrative in which the child shows a deficit.[17] Gillam and colleagues reported that the total MISL score had acceptable inter-rater reliability, internal consistency reliability, and construct validity for specific aspects of the narrative.[17] Therefore, based on the MISL scores, we confirmed the validity and reliability of our test method.

The results of the construct validity showed strong correlations between each MISL item and the total score of each test (r≥0.5). Moreover, it was confirmed that all MISL items were essential for constructing a story, had an effect on the NC score, and the pictures used in our study were related to all items. However, some items had a significant effect on the NC score (e.g., character, setting, initial event, and conclusion), whereas microstructure items had less effect. The observed difference could be associated with younger age (6-8 years old) of the participants in our study. A previous study on the narrative development in TD children reported that children at the age of 9 can construct a complex story. They use temporal and causal connections to describe the mental state and emotions of the depicted characters.[30] In comparison with the age group in our study, 9-year-olds read more books and have more experience in constructing a story, which in turn affects the NC score through microstructure items.

A comparison of the NC score between the children in the TD and ASD groups confirmed the discriminant validity, since our tests could discriminate between these groups. Our findings were in line with the results of previous studies on autistic children.[12, 26, 31] Internal consistency represents the homogeneity of items that have been considered to measure a specific construct. Our results showed acceptable levels of internal consistency for all pictures. However, in line with a previous study,[17] Cronbach's alpha of the total MISL score was higher than that of the individual sub-scales. The results showed that the total MISL score was a more reliable indicator of change during the intervention than the individual macrostructure or microstructure scores. However, this does not contradict the use of an individual subscale's score to monitor progress as well as deciding on specific objectives for intervention sessions. The test-retest reliability showed that the ICC value between the two

interventions was significant for all stories (>0.70). Stockman[32] found that measuring language skills at two different times cannot be reliable due to the context effect. Another study stated that an increase in the length of language samples increased reliability.[30] Although the language samples used in our study were short for all stories (17 phrases on average), the ICC values were high. Two factors could be the reasons for these high ICC values. First, due to the difference in variables used in the analysis of narratives compared to other studies. Second, the number of total words, language structures, and personal pronouns were investigated more extensively in previous studies,[4, 5] while these items were unstable in our study. Therefore, instead, we used the MISL to investigate the main narrative subscales (macrostructures and microstructures). In addition, we conducted an exercise session before the actual tests to familiarize the children with the procedure. Moreover, the tests were repeated after one month, which contributed to achieving high ICC values.

We identified two main limitations in our study. First, the transcription and analysis of language samples were time-consuming, which forced us to reduce the number of participants. Second, since the tests were repeated with a 1-month interval, the data of those participants absent in any of these tests were omitted.

Conclusion

The developed tests can be used to evaluate a variety of narrative skills. The results showed that the designed pictures stimulated the quality of story retelling ability in children. The tests can be used by clinicians and researchers to diagnose narrative deficit and design appropriate interventions. It is recommended that future studies include more typically developing children and cover a broader range of language impairments to assess narrative skills in children and to substantiate our findings.

Acknowledgment

The present manuscript was extracted from the PhD thesis of Saeide Beytollahi. The study was not financially supported by any organization. The authors would like to thank Mohamad Mohseni for his assistance and contribution to the study. Finally, special thanks are due to the children and their parents for their time, effort, and patience.

Conflict of Interest: None declared.

References

1 Spencer TD, Slocum TA. The effect of a narrative intervention on story retelling and personal story generation skills of preschoolers with risk factors and narrative language delays. Journal of Early Intervention. 2010;32:178-99. doi: 10.1177/1053815110379124.

2 Costa GM, Rossi NF, Giacheti CM. Performance of Brazilian Portuguese speakers in the Test of Narrative Language (TNL). Codas. 2018;30:e20170148. doi: 10.1590/2317-1782/20182017148. PubMed PMID: 30043829.

3 Mehta PD, Foorman BR, Branum-Martin L, Taylor WP. Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. Scientific Studies of Reading. 2005;9:85-116. doi: 1207/s1532799xssr0902_1.

4 Pankratz ME, Plante E, Vance R, Insalaco DM. The diagnostic and predictive validity of the Renfrew Bus Story. Lang Speech Hear Serv Sch. 2007;38:390-9. doi: 10.1044/0161-1461(2007/040). PubMed PMID: 17890518.

5 Heilmann J, Miller JF, Nockerts A, Dunaway C. Properties of the narrative scoring scheme using narrative retells in young school-age children. Am J Speech Lang Pathol. 2010;19:154-66. doi: 10.1044/1058-0360(2009/08-0024). PubMed PMID: 20008470.

6 Petersen DB, Spencer TD. The narrative language measures: Tools for language screening, progress monitoring, and intervention planning. Perspectives on Language Learning and Education. 2012;19:119-29. doi: 10.1044/lle19.4.119.

7 Ketelaars MP, Jansonius K, Cuperus J, Verhoeven L. Narrative competence in children with pragmatic language impairment: a longitudinal study. Int J Lang Commun Disord. 2016;51:162-73. doi: 10.1111/1460-6984.12195. PubMed PMID: 26935766.

8 Soodla P, Kikas E. Macrostructure in the narratives of Estonian children with typical development and language impairment. J Speech Lang Hear Res. 2010;53:1321-33. doi: 10.1044/1092-4388(2010/08-0113). PubMed PMID: 20643787.

9 Eisenberg SL, Ukrainetz TA, Hsu JR, Kaderavek JN, Justice LM, Gillam RB. Noun phrase elaboration in children's spoken stories. Lang Speech Hear Serv Sch. 2008;39:145-57. doi: 10.1044/0161-1461(2008/014). PubMed PMID: 18420517.

10 Nippold MA. Later language development: School-age children, adolescents, and young

adults. California: ERIC; 2016.
11 Wong AM-Y, Kidd JC, Ho CS-H, Au TK-F. Characterizing the overlap between SLI and dyslexia in Chinese: The role of phonology and beyond. Scientific Studies of Reading. 2010;14:30-57. doi: 10.1080/10888430903242043.
12 Gillam SL, Hartzheim D, Studenka B, Simonsmeier V, Gillam R. Narrative Intervention for Children With Autism Spectrum Disorder (ASD). J Speech Lang Hear Res. 2015;58:920-33. doi: 10.1044/2015_JSLHR-L-14-0295. PubMed PMID: 25837265.
13 Lok B, McNaught C, Young K. Criterion-referenced and norm-referenced assessments: compatibility and complementarity. Assessment & Evaluation in Higher Education. 2016;41:450-65. doi: 10.1080/02602938.2015.1022136.
14 Reilly J, Losh M, Bellugi U, Wulfeck B. "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. Brain Lang. 2004;88:229-47. doi: 10.1016/S0093-934X(03)00101-9. PubMed PMID: 14965544.
15 Soleymani Z, Nematzadeh S, Tehrani LG, Rahgozar M, Schneider P. Language sample analysis: development of a valid language assessment tool and determining the reliability of outcome measures for Farsi-speaking children. European Journal of Developmental Psychology. 2016;13:275-91. doi: 10.1080/17405629.2016.1152175.
16 Petersen DB, Gillam SL, Gillam RB. Emerging procedures in narrative assessment: The index of narrative complexity. Topics in language disorders. 2008;28:115-30. doi: 10.1097/01.TLD.0000318933.46925.86.
17 Gillam SL, Gillam RB, Fargo JD, Olszewski A, Segura H. Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills. Communication Disorders Quarterly. 2017;38:96-106. doi: 10.1177/1525740116651442.
18 Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34-42. doi: 10.1016/j.jclinepi.2006.03.012. PubMed PMID: 17161752.
19 Grant R, Nozyce M. Proposed changes to the American Psychiatric Association diagnostic criteria for autism spectrum disorder: implications for young children and their families. Matern Child Health J. 2013;17:586-92. doi: 10.1007/s10995-013-1250-9. PubMed PMID: 23456348.
20 Ahmadi S, Hemmatian S, Khalili Z. Investigation of the psychometric features of the GARS (persion). Research in Cognitive and Behavioral Journal. 2011;1:87-104.
21 Samadi SA, Mahmoodizadeh A, McConkey R. A national study of the prevalence of autism among five-year-old children in Iran. Autism. 2012;16:5-14. doi: 10.1177/1362361311407091. PubMed PMID: 21610190.
22 Razavieh A, Shahim S. Retest reliability of the Wechsler Preschool and Primary Scale of Intelligence restandardized in Iran. Psychol Rep. 1990;66:865-6. doi: 10.2466/pr0.1990.66.3.865. PubMed PMID: 2377704.
23 Shurman J, Leone D. Standardized Versus Naturalized: An Evaluation of Child Morphological and Syntactic Assessments. International Journal of Undergraduate Research and Creative Activities. 2015;7:6. doi: 10.7710/2168-0620.1064.
24 Tang W, Hu J, Zhang H, Wu P, He H. Kappa coefficient: a popular measure of rater agreement. Shanghai Arch Psychiatry. 2015;27:62-7. doi: 10.11919/j.issn.1002-0829.215010. PubMed PMID: 25852260; PubMed Central PMCID: PMCPMC4372765.
25 Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the academy of marketing science. 2015;43:115-35. doi: 10.1007/s11747-014-0403-8.
26 Eigsti I-M, de Marchena AB, Schuh JM, Kelley E. Language acquisition in autism spectrum disorders: A developmental review. Research in Autism Spectrum Disorders. 2011;5:681-91. doi: 10.1016/j.rasd.2010.09.001.
27 Cho E, Kim S. Cronbach's coefficient alpha: Well known but poorly understood. Organizational Research Methods. 2015;18:207-30. doi: 10.1177/1094428114555994.
28 Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19:231-40. doi: 10.1519/15184.1. PubMed PMID: 15705040.
29 Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977;33:363-74. doi: 10.2307/2529786. PubMed PMID: 884196.
30 Westby C, Culatta B. Telling Tales: Personal Event Narratives and Life Stories. Lang Speech Hear Serv Sch. 2016;47:260-82. doi: 10.1044/2016_LSHSS-15-0073. PubMed PMID: 27679843.

31 Siller M, Swanson MR, Serlin G, George A. Internal State Language in the Storybook Narratives of Children with and without Autism Spectrum Disorder: Investigating Relations to Theory of Mind Abilities. Res Autism Spectr Disord. 2014;8:589-96. doi: 10.1016/j.rasd.2014.02.002. PubMed PMID: 24748899; PubMed Central PMCID: PMCPMC3988534.

32 Stockman IJ. The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. Language, Speech, and Hearing Services in Schools. 1996;27:355-66. doi: 10.1044/0161-1461.2704.355.