International Journal of Pediatric Otorhinolaryngology 89 (2016) 17–24

Contents lists available at ScienceDirect

International Journal of Pediatric Otorhinolaryngology

journal homepage: http://www.ijporlonline.com/

Reliability, validity and normative data of the Italian version of the Bus Story test

F. Mozzanica [a,*], R. Salvadorini [b], E. Sai [c], R. Pozzoli [d], P. Maruzzi [a], L. Scarponi [a], M.R. Barillari [e], E. Spada [f], F. Ambrogi [f], A. Schindler [a]

[a] Department of Biochemical and Clinical Science "Luigi Sacco", University of Milan, Milan, Italy
[b] IRCCS, Calambrone, Italy
[c] AO Poma, Mantova, Italy
[d] IRCCS Medea La Nostra Famiglia, Bosisio Parini, Italy
[e] Phoniatric Department, University of Naples, Naples, Italy
[f] Clinical Sciences and Community Department, University of Milan, Milan, Italy

ARTICLE INFO

Article history:
Received 17 January 2016
Received in revised form
17 July 2016
Accepted 21 July 2016
Available online 25 July 2016

Keywords:
Bus Story test
Psychometrics
Validity
Story retelling

ABSTRACT

Objectives: Evaluation of the reliability and the validity of the Italian version of the Bus Story Test (I-BST), providing normative data in Italian children.
Methods: A total of 552 normally developing children (278 males and 274 females) aged 3; 6 to 9; 0 years, were enrolled. Test-retest, intra- and inter-rater reliability were analysed on a sample of respectively 145, 178 and 178 children. Normative data were gathered from all the enrolled children and estimate centiles according to the CG-LMS method provided. The children were divided into 11 age classes of six months each; percentile scores and standard error measurement were analysed in children from age class 4; 0-4; 5 years to age class 8; 6-811 years. Age effects on I-BST were analysed.
Results: Results showed high test-retest, intra- and inter-rater reliability scores. A significant age effect on I-BST scores emerged from the ANOVA test analysis; in particular, as age increases, so do I-BST scores.
Conclusion: The I-BST is a reliable and valid tool. The availability of normative data for Italian speaking children may help clinicians during clinical assessment.

1. Introduction

An oral narrative can be defined as a monologue describing an experience or events that are chronologically sequenced. This is a demanding task, and it relies on genre-specific content (for example the conventional forms used as introductions and closings), structural knowledge (plot development based on events linked through causal relationships), and linguistic knowledge [1]. In particular, fictional narratives involve the ability to verbally express text units, to be detached from what is happening, and to build a coherent plot covering several events placed in the right temporal order and linked by the most appropriate causal relationship. Linguistically, the narrator must be able to encode information about the characters and events of the story, and to match them according to a temporal logic [2]. Cognitively, the narrator must be able to understand and express the core elements, the actions, the logical relationships between events, and the theme of the story. Finally, in order to make oneself successfully understood, the narrator also needs to take into account the interlocutor's needs, reactions, and motivations [3].

Narration is crucial in carrying out everyday activities, such as communicating one's own personal experiences and is a crucial part of a child's daily life both at home and in school [4]. In addition, narration has been found to be a valid predictor of long-term language skills and to play an important role in academic achievement and social success [5]. Several studies found narration to be a good predictor of later reading-comprehension, reading-fluency, or written-narrative skills in both children who have learning disabilities and in those who are typically developing [6–8]. Furthermore, narrative skills uniquely contribute to reading fluency even after controlling for receptive vocabulary and decoding skills [9]. For this reason, narrative ability analysis is considered one of the most interesting and contextually valid methods to measure

* Corresponding author. Department of Biochemical and Clinical science "Luigi Sacco", Via G.B. Grassi 74, Milan, 20100, Italy.
E-mail address: francesco.mozzanica@gmail.com (F. Mozzanica).

http://dx.doi.org/10.1016/j.ijporl.2016.07.029
0165-5876/

communication competence, both in healthy and pathological paediatric populations [5]. Besides, it is a diagnostic criterion for Language Disorder in the DSM-5 [10].

Impairment in narrative abilities is common to different clinical conditions often found in paediatric otorhinolaryngology. In particular, specific language impairment (SLI) seems to affect narrative skills, since children with SLI produce stories that are shortened, in comparison to their typically developing age-matched peers [11–13], and are similar to those of younger children with typical language development. In addition, even after other language skills such as semantics and syntax improve, children with SLI may continue to exhibit a deficiency in narrative abilities [14].

Narrative difficulties have been reported in children affected by, for example, verbal comprehension difficulties [5], cerebral palsy [15], deafness [16], or hydrocephalus [17].

The evaluation of narrative ability can be performed through the administration of a specific narrative task, either in the form of a story generation task or as a story retelling task, and taking into consideration two levels – the micro- and the macro-level [12,18]. On one hand, the micro-level often focuses on the diversity of words and the frequency and complexity of sentences employed. On the other hand, the macro-level evaluates the ability to organize the elements of the narrative in proper grammatical structures and, in the case of story retelling tasks, it also evaluates the number of information units provided by the subject. The form of the task appears to be crucial, since young children have been found to be sensitive to different elicitation tasks and story genres [1,4]. Story generation is considered more complicated, since it originates from children themselves without the intervention of external stimuli [19]. Furthermore, it results more representative of spontaneous communication, and reflects the natural form of discourse. Story retelling tasks are generally less demanding, and for this reason they appear particularly appropriate to test preschool children. Besides, the clinician is familiar with the content of the story, thus making the scoring easier and more reliable [18].

Even if the impairment of narrative abilities may have important clinical consequences, there are only a few specific narrative tasks with proven reliability and clinical validity available. In particular, the Bus Story Test (BST) [20] is a norm-referenced measure of young children's narrative abilities suitable for ages ranging from preschool to kindergarten. This assessment tool is simple, attractive, easily administered, and was found to predict persistent language impairment and to have strong relationship to later literacy [21].

The BST was originally developed in England, and an adaptation for American children was issued in 1994 [22]. The English version was normed in 1993–1994 on 573 children from South-East England, ages 3.6–8.0 years. Reliability was tested on 13 children, but no formal statistics were reported in the manual [23]. The American adaptation was normed on a population of 418 schoolchildren, ages 3.0–6.11 years, from both urban and rural settings. Children with hearing impairments, language delays, learning disabilities, or otherwise identified by the teacher or school as non-typically developing were excluded. Test-retest reliability was tested on 27 children, ages 4.0–6.11 years, and scores for Information, Sentence Length, and Complexity were $r = 0.79$, $r = 0.72$, and $r = 0.58$ respectively. Inter-rater reliability was analysed comparing the BST scores obtained from 25 transcripts analysed both by two teachers and by two authors. The correlation scores between the two teachers for Information, Sentence Length and Complexity were 0.92, 0.70, and 0.22 respectively. The correlation scores between the two teachers' scores and the authors' were 0.72 and 0.70 for Information; 0.83 and 0.81 for Sentence Length and 0.60 and 0.33 for Complexity. In that study, children performed increasingly well as a function of chronological age, but no statistical analysis of variation across age was reported. The BST has been widely used in research focused on children who are developing their language in a typical fashion, as well as on those with language impairments [24–28].

An Italian version of the BST (I-BST) has already been developed, and pilot-tested on a sample of 80 typically developing preschoolers [29]. However, the psychometric characteristics of the I-BST, including its reliability and validity, were not analysed and no information regarding the narrative abilities of school-aged, Italian-speaking children are available. The availability of reliable and valid I-BST will allow the clinical assessment of narrative abilities in Italian-speaking children. In addition, the availability of normative data on normally developing children will help clinicians in the evaluation of their narrative abilities in both normal and pathological children.

The aim of this study is to evaluate the reliability and the validity of the I-BST [29], and provide normative data.

2. Material and methods

The study consisted of three different phases: reliability analysis (phase 1), normative data generation (phase 2), validity analysis (phase 3). All data were collected prospectively. Parents or guardians provided written informed consent for each subject enrolled. Recruited children, parents, and guardians involved in the project were clearly informed, and agreed to participate without any compensation. The study was carried out in compliance with the Declaration of Helsinki.

2.1. I-BST

The I-BST was used for narrative abilities assessment. The cross-cultural adaptation of this test had been previously performed using standard techniques [29,30]. In the study of Zarmati et al. [29], each item of the original test was translated into Italian by a professional translator and two bilingual investigators. Two independent phoniatricians (medical doctors who underwent a 5-year-long residency focusing on voice, speech, language, hearing, and swallowing disorders) familiar with the process of instrument validation, examined semantics, idiomatic and conceptual issues, and therefore were able to further refine these versions. An Italian final-consensus version was obtained and given to two professional translators, who were asked to translate the test literally back into English. Once this task was completed, the two translators and an expert committee synthesized the results of this back-translation, which was then compared with the original English version of the test to check that they actually kept the same semantic value.

The BST includes a story about a bus that runs away from its driver with twelve accompanying pictures. During the assessment, the rater starts telling the story, and then asks the child to retell the same story using the pictures in a wordless storybook as prompts. The story retold by the child is audio and video recorded, transcribed, and scored on the macro- and micro-level measures described in the test manual. Among these measures, the Information subscale is a macro-level measure, and it indicates how many of the 32 key-information units of the original story the child uses while retelling the story (the child consequently can get credit for a response that matches the key-information even if he uses different words). The total possible raw score is 52 since it includes some items worth 2 points. This subscale reflects the child's proficiency level on a set of integrated skills (i.e., memory, vocabulary, story knowledge). On the other hand, Sentence Length and Complexity are the micro-level measures included to indicate, respectively, morpho-syntactic complexity (calculated as the

average number of words in the five longest sentences that a child generates in his or her retell) and syntactic development (calculated as the number of utterances containing a subordinate clause).

2.2. Participants

For the purposes of this study, a population of 552 typically developing children were recruited. All 552 typically developing children, 278 males and 274 females, ages ranging from 36 to 101 months, were Italian coming from a wide range of backgrounds, including children with a low socio-economic status. Exclusion criteria were: intellectual disability, deafness, bilingualism, cerebral palsy and any other motor impairment, speech-organs impairment of any origin, such as cleft palate, and language impairment of any origin. Normally developing children were recruited in both urban and non-urban classrooms of public kindergartens and schools in Northern Italy. All information on exclusion criteria were obtained by both teachers and parents, who filled out a questionnaire, specifically designed for this study, asking for the presence of any of the above-mentioned exclusion criteria.

2.3. Assessment procedure

Children were individually administered the I-BST by trained assessors. A total of 8 assessors were enrolled, and they all completed a 4-h training program specific to I-BST administration. This training program addressed both I-BST test administration and result analysis. All assessors were speech-language pathologists (SLPs) from our hospital, who had at least 5 years of experience in child language assessment. Each SLP rated children from different age groups. All assessments were video and audio recorded in order to facilitate transcription. The administration of the test never took longer than 10 min. The transcription process was performed by the same SLP who administered the test within 12 h from it. The results analysis was performed immediately after the transcription by the same SLP who managed to evaluate the Information, Sentence Length, and Complexity scores. The transcription process and the results analysis never exceeded 20 min in total.

2.4. Reliability analysis

One hundred forty-five children out of 552 typically developing children were randomly selected for reliability analysis. For test-retest reliability, the same rater administered the I-BST twice within a span of approximately two weeks. The length of this interval was selected because no substantial change was expected to take place in children's narrative abilities over this period. The results analysis was performed by the same SLP who carried out the testing, within 12 h from test administration. During the second administration, the rater did not have any access to the scores obtained during the first evaluation. Test-retest reliability was assessed through two-way mixed-effects model (consistency definition) intraclass correlation coefficients (ICCs), both for the macro- and micro-level measures, as described in the test manual. In order to evaluate I-BST intra- and inter-rater reliability, a random sample of 178 recordings was listened to and rated by two licensed speech-language pathologists from the same facility, named rater 1 and 2, specialized in the assessment and management of language impairment. The two raters managed to complete this task twice, with a week of interval in order to evaluate also the intra-rater reliability. Both intra- and inter-rater reliability were evaluated with ICCs.

2.5. Normative data generation

The scores obtained in the I-BST test by each of the 552 typically developing children were included to establish normative data. Like in the English manual [23], the cohort of subjects was divided into 11 different groups according to their ages (Table 1). Each of the 11 groups encompassed a 6-month age range. Only 9 out of the 11 age groups were composed by at least 50 children; consequently only these 9 age groups were enrolled for normative data generation.

In order to provide information about the subjects' socio-economic-status (SES) background, children's parents were asked to complete a brief questionnaire regarding their current employment status and education level, while the I-BST was taking place with their children. Educational levels and parental type of employment were organized into three categories: both educational and parental work environment were classified as high, low, or mixed (different combinations of high, medium, and low). In order to proceed to such categorization, educational levels were first classified as low (<=8 years), medium (9–13 years), or high (>=14) while parents' employment status was classified into 4 different groups, i.e., unemployed, manual workers, office workers, and intellectual workers. Families with both parents unemployed were grouped with families whose parents were both manual workers.

2.6. Validity

For construct validity analysis, the I-BST test's macro- and micro-level scores were compared across the 9 age groups with at least 50 children.

2.7. Statistical analysis

Statistical tests were performed using the SPSS 19.0 statistical software (SPSS, Inc., Chicago, IL). One of the goals was a deeper analysis of the reliability and validity of I-BST compared to the original and the American studies. For this reason, a larger number of children were recruited for reliability analysis and a different type of validity were analysed in comparison to the original psychometric studies. ICC was used to evaluate I-BST's test-retest, inter-rater and intra-rater reliability. To estimate the centiles, the CG-LMS method was used on all the 552 recruited children [31]. The CG-LMS method is a model that expresses the centiles in terms of age-specific curves called L, M, and S. The M and S curves correspond to the median and coefficient of variation of Information,

Table 1
Population recruited for the study. Depending on the age of the participants 11 different categories were considered. The number of typical developing children within each category and their gender are reported.

Age range (years; months)	Number typical developing		
	Males	Females	Total
3; 6-3; 11	6	8	14
4; 0-4; 5	24	26	50
4; 6-4; 11	25	26	51
5; 0-5; 5	27	27	54
5; 6-5; 11	29	30	59
6; 0-6; 5	30	32	62
6; 6-6; 11	30	37	67
7; 0-7; 5	34	26	60
7; 6-7; 11	32	24	56
8; 0-8; 5	26	27	53
8; 6-8; 11	15	11	26
	278	274	552

Complexity and Sentence Length at each age, whereas the L curve allows for the age-dependent skewness of the distribution of the same trait. For the complexity score a Poisson distribution was used and results should be taken with caution. For each age class including at least 50 children mean, median, standard deviation, 1st and 3rd quartile, 10th and 90th percentile as well as standard error of measurement were calculated. In these children ANOVA test with Tukey post-hoc was employed in order to evaluate age effects on I-BTS's macro- and micro-level scores. Due to the large number of comparisons, a more stringent level of significance was set using Bonferroni correction. A Chi-Square test was used to evaluate the differences in maternal and paternal employment and educational levels among the enrolled children. The significance level was set at 0.05 across all statistical analyses with an exception for ANOVA's comparisons, for which the significance level was set at 0.005, after Bonferroni correction.

3. Results

No difficulty emerged from the administration of the I-BST to all 552 children enrolled in the study by the chosen trained speech-language pathologists.

3.1. Reliability analysis

The I-BST macro- and micro-level-measure scores obtained for test-retest, intra- and inter-rater reliability analysis are reported in Table 2. Test-retest, intra-rater and inter-rater reliabilities for each of the 3 scales were all at least r = 0.90.

3.2. Normative data

Estimate of the I-BST Information, Sentence Length and Complexity measures centiles according to the CG-LMS method are provided respectively in Figs. 1–3. For the 3 different I-BUS Story test measures an evolution of percentiles with age is clearly visible.

The results obtained during the narrative assessment of the 9 different age groups including at least 50 children per age class are reported in Tables 3–5. For the Information measure, the number of key-information units of the story the child uses in his story retell increased with the age. Morpho-syntactic complexity and syntactic development also increased with age. When we analysed parents' SES, manual work was the most common employment among fathers, while office work was the most common employment among mothers (Chi-Square test p-value = 0.001). Education level was slightly higher for mothers (Chi-Square test p-value = 0.001). Parents with a mixed level in both educational- and work-environment composed the large majority of the families enrolled. Only 21 families were composed by both unemployed/manual worker parents with low education levels. On the other hand, only 43 families were composed by both intellectual worker parents with high education levels.

3.3. Validity analysis

The number of key-information units of the original story that the child uses in his/her story retell increases with age. In typically developing children, the ANOVA test showed a significant age effect for the I-BST Information-measure score [F (10, 501) = 50.4, p = 0.001]. Results of post-hoc analysis with Tukey test are reported in Table 6. From the age class 4;0-4;5 to the age class 5;6-5;11 statistically significant difference in the Information measure scores were found when the compared age-classes were spaced-out by 1 year at least; however, no statistically significant differences in the Information measure scores were found between the age class 6;0-6;5 with respect to 7;0-7;5 (p = 0.131), and the age class 6;6-6;11 with respect to 7;6-7,11 (p = 0.486).

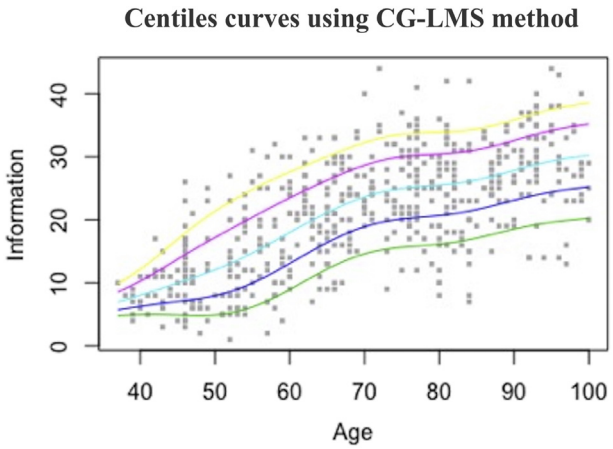

Fig. 1. Estimate of the I-BUS Story test Information measures centiles according to the CG-LMS method. The centiles in terms of age-specific curves are expressed. Green, dark blue, light blue, purple and yellow curves respectively represent 10°, 25°, 50°, 75° and 90° percentiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

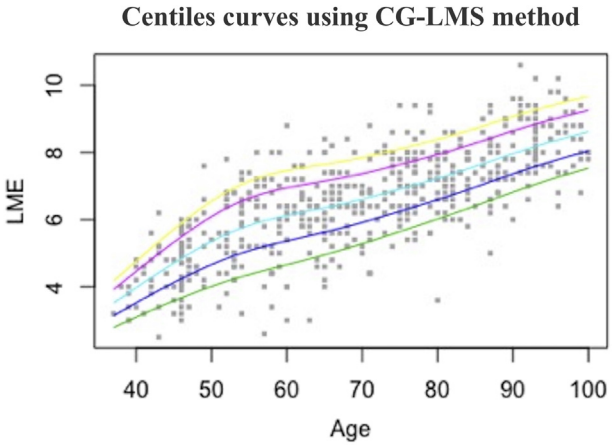

Fig. 2. Estimate of the I-BUS Story test Sentence Length (LME) measures centiles according to the CG-LMS method. The centiles in terms of age-specific curves are expressed. Green, dark blue, light blue, purple and yellow curves respectively represent 10°, 25°, 50°, 75° and 90° percentiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Reliability analysis of the I-BST. The ICC values for test-retest, intra-rater and inter-rater reliability of the I-BST are reported. Ranges are reported in brackets.

	ICC Test-retest (n = 145)	ICC Intrarater reliability (n = 178)	ICC Interrater reliability (n = 178)
Information	0.93 (0.86–0.95)	0.98 (0.92–0.99)	0.92 (0.82–0.96)
Sentence length	0.98 (0.93–0.99)	0.96 (0.86–0.99)	0.91 (0.72–0.96)
Complexity	0.80 (0.75–0.84)	0.95 (0.87–0.98)	0.90 (0.81–0.94)

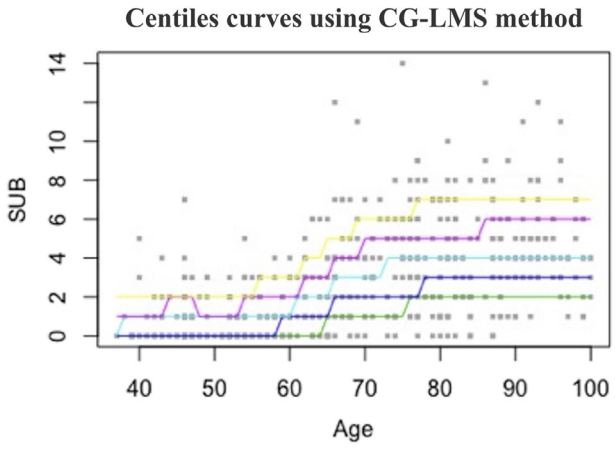

Fig. 3. Estimate of the I-BUS Story test Complexity (SUB) measures centiles according to the CG-LMS method. The centiles in terms of age-specific curves are expressed. Green, dark blue, light blue, purple and yellow curves respectively represent 10°, 25°, 50°, 75° and 90° percentiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Also in the Sentence length measure the ANOVA test, with Tukey post-hoc demonstrated a significant effect for age [$F(10, 501) = 45.7$, $p = 0.001$]. Results of post-hoc analysis with Tukey test are reported in Table 7. From the age class 4;0-4;5 to the age class 6;0-6;5 statistically significant difference in the Sentence lenght measure scores were found when the compared age-classes were spaced-out by 1 year at least. No statistically significant differences in the Sentence lenght measure scores were found between the age class 6;6-6;11 with respect to 7;6-7;11 ($p = 0.280$), and the age class 7;0-7;5 with respect to 8;0-8,5 ($p = 0.360$).

Finally, the ANOVA test demonstrated a significant effect of age for the Complexity measure score of the I-BST [$F(10, 501) = 25.7$, $p = 0.001$]. Results of post-hoc analysis with Tukey test are reported in Table 8. From the age class 4;0-4;5 to the age class 6;6-6;11 statistically significant difference in the Complexity measure scores were found when the compared age-classes were spaced-out by 1 year at least. No statistically significant differences in the Complexity measure scores was found between the age class 7;0-7;5 with respect to 8;0-8;5 ($p = 0.340$).

Table 3
Information (INFO) scores obtained in the 9 age categories of typical developing children. Mean, standard deviation (SD), median, 1st, 3rd quartiles, 10th, 50th, 90th percentiles and standard error (SE) are reported.

Age range (years; months)	Mean	Median	SD	1st quartile	3rd quartile	10th percentile	50th percentile	90th percentile	SE
4; 0-4; 5	10.50	11.00	4.30	7.25	12.00	4.90	11.00	16.10	0.61
4; 6-4; 11	11.92	11.00	6.48	6.00	16.00	5.00	11.00	20.00	0.91
5; 0-5; 5	15.76	15.00	7.50	10.00	20.00	7.00	15.38	25.70	1.02
5; 6-5; 11	20.17	21.00	6.33	15.50	24.50	11.00	21.00	27.40	0.82
6; 0-6; 5	23.55	24.00	6.05	20.00	27.75	16.00	24.00	30.00	0.77
6; 6-6; 11	24.48	24.00	5.83	21.00	28.50	18.00	24.00	32.40	0.71
7; 0-7; 5	24.92	25.00	6.21	20.50	29.00	16.00	25.00	32.10	0.80
7; 6-7; 11	25.30	26.00	6.76	20.00	30.25	17.00	26.00	32.50	0.90
8; 0-8; 5	30.40	31.00	6.47	27.00	35.00	23.20	31.00	37.80	0.89

Table 4
Sentence Length scores obtained in the 9 age categories of typical developing children. Mean, standard deviation (SD), median, 1st, 3rd quartiles, 10th, 50th, 90th percentiles and standard error (SE) are reported.

Age range (years; months)	Mean	Median	SD	1st quartile	3rd quartile	10th percentile	50th percentile	90th percentile	SE
4; 0-4; 5	4.62	4.80	0.91	4.00	5.28	3.40	4.71	5.80	0.13
4; 6-4; 11	5.34	5.20	0.99	4.60	5.90	4.20	5.20	6.80	0.14
5; 0-5; 5	5.97	6.20	1.22	5.20	6.80	4.40	6.20	7.20	0.17
5; 6-5; 11	6.15	6.20	1.22	5.70	6.80	4.52	6.20	7.60	0.16
6; 0-6; 5	6.41	6.45	0.98	5.65	7.00	5.20	6.43	7.40	0.12
6; 6-6; 11	7.06	7.00	0.93	6.40	7.80	6.00	7.00	8.00	0.11
7; 0-7; 5	7.21	7.20	0.96	6.60	7.85	6.18	7.20	8.42	0.12
7; 6-7; 11	7.57	7.60	0.94	6.95	8.05	6.40	7.60	8.90	0.13
8; 0-8; 5	8.40	8.40	0.83	7.80	9.00	7.40	8.40	9.40	0.11

Table 5
Complexity scores obtained in the 9 age categories of typical developing children. Mean, standard deviation (SD), median, 1st, 3rd quartiles, 10th, 50th, 90th percentiles and standard error (SE) are reported.

Age range (years; months)	Mean	Median	SD	1st quartile	3rd quartile	10th percentile	50th percentile	90th percentile	SE
4; 0-4; 5	0.98	0.00	1.63	0.00	1.00	0.00	0.00	2.00	0.23
4; 6-4; 11	0.88	1.00	0.99	0.00	2.00	0.00	1.00	2.00	0.14
5; 0-5; 5	1.07	1.00	1.23	0.00	2.00	0.00	1.00	3.00	0.17
5; 6-5; 11	1.90	1.00	1.74	1.00	3.00	0.00	1.00	4.00	0.23
6; 0-6; 5	3.35	3.00	2.38	2.00	4.00	1.00	3.00	6.00	0.30
6; 6-6; 11	4.06	4.00	2.60	2.00	6.00	1.00	4.00	7.00	0.32
7; 0-7; 5	3.83	4.00	2.51	2.00	6.00	1.00	4.00	7.00	0.32
7; 6-7; 11	3.98	4.00	2.75	2.00	6.00	1.00	4.00	8.00	0.37
8; 0-8; 5	4.92	5.00	1.98	4.00	6.00	3.00	5.00	8.00	0.27

Table 6
Results of post-hoc comparison of the Information level scores with Tukey test are reported.

Age range (years; months)	4; 0-4; 5	4; 6-4; 11	5; 0-5; 5	5; 6-5; 11	6; 0-6; 5	6; 6-6; 11	7; 0-7; 5	7; 6-7; 11	8; 0-8; 5
4; 0-4; 5	–	0.974	0.005	0.000	0.000	0.000	0.000	0.000	0.000
4; 6-4; 11		–	0.493	0.000	0.000	0.000	0.000	0.000	0.000
5; 0-5; 5			–	0.625	0.000	0.000	0.000	0.000	0.000
5; 6-5; 11				–	0.453	0.021	0.041	0.001	0.000
6; 0-6; 5					–	0.258	0.131	0.018	0.001
6; 6-6; 11						–	0.645	0.486	0.004
7; 0-7; 5							–	0.745	0.039
7; 6-7; 11								–	0.184
8; 0-8; 5									–

Table 7
Results of post-hoc comparison of the Sentence lenght level scores with Tukey test are reported.

Age range (years; months)	4; 0-4; 5	4; 6-4; 11	5; 0-5; 5	5; 6-5; 11	6; 0-6; 5	6; 6-6; 11	7; 0-7; 5	7; 6-7; 11	8; 0-8; 5
4; 0-4; 5	–	0.770	0.010	0.000	0.000	0.000	0.000	0.000	0.000
4; 6-4; 11		–	0.460	0.003	0.000	0.000	0.000	0.000	0.000
5; 0-5; 5			–	0.360	0.010	0.000	0.000	0.000	0.000
5; 6-5; 11				–	0.650	0.030	0.041	0.000	0.000
6; 0-6; 5					–	0.540	0.010	0.005	0.001
6; 6-6; 11						–	0.130	0.280	0.001
7; 0-7; 5							–	0.685	0.360
7; 6-7; 11								–	0.585
8; 0-8; 5									–

Table 8
Results of post-hoc comparison of the Complexity level scores with Tukey test are reported.

Age range (years; months)	4; 0-4; 5	4; 6-4; 11	5; 0-5; 5	5; 6-5; 11	6; 0-6; 5	6; 6-6; 11	7; 0-7; 5	7; 6-7; 11	8; 0-8; 5
4; 0-4; 5	–	0.760	0.003	0.000	0.000	0.000	0.000	0.000	0.000
4; 6-4; 11		–	0.540	0.001	0.000	0.000	0.000	0.000	0.000
5; 0-5; 5			–	0.430	0.020	0.000	0.000	0.000	0.000
5; 6-5; 11				–	0.530	0.020	0.041	0.000	0.000
6; 0-6; 5					–	0.660	0.030	0.001	0.000
6; 6-6; 11						–	0.230	0.040	0.001
7; 0-7; 5							–	0.350	0.340
7; 6-7; 11								–	0.435
8; 0-8; 5									–

4. Discussion

The BST is a norm-referenced measure of young children's narrative abilities used along with other assessment tools to guide further areas of diagnostic testing. In fact, the BST measures children's ability to retell relevant narrative concepts about a story, thus providing information about the children's integrative language skills using a naturally occurring activity [32].

The BST has been widely used in research focused on children who are developing their language in a typical fashion as well as on those with language impairments [24–28]. In particular, Pankratz et al. [33], who studied the predictive validity of the BST by comparing the clinical results obtained in normal and SLI children, reported that the BST could be used as an indicator of future language performance for children with SLI.

Although an Italian version of the BST had been already developed and pilot-tested on a small sample of typically developing children before this study [29], its psychometric characteristics were lacking. In addition, the currently available norm-referenced data had previously been obtained only from English-speaking children. In the present study, I-BST's psychometric properties and normative data were studied in an Italian population. Results showed good test-retest, intra- and inter-rater reliability, and criterion validity. Therefore, these results further support I-BST's application as a reliable tool for narrative assessment.

Some I-BST-specific findings are noteworthy. In particular, I-BST was administered to all enrolled children in less than 10 min per subject. Consequently, I-BST might be speculated not to be a burdensome instrument, and to be easily administered.

As far as I-BST test-retest reliability is concerned, the scores obtained in test-retest analysis support the idea that I-BST has a high stability and reproducibility over time. In fact, ICC scores were $r = 0.93$, $r = 0.98$, and $r = 0.80$ for Information, Sentence Length, and Complexity, respectively. In the American study, test-retest correlation coefficients were $r = 0.79$ for Information, 0.72 for Sentence Length, and 0.58 for Complexity [22]. It is possible that these differences are related to the cohort of patients enrolled for test-retest reliability analysis. In fact, in the American study, only 27 children were tested twice with a one- or two-month-long interval, while in the present study a total of 145 children were tested twice by the same rater within a 2-week-long interval.

I-BST intra- and inter-rater reliability appeared to be satisfactory since all values were at least $r = 0.90$. To the best of our knowledge, no other data regarding BST intra-rater reliability are available. Only little information is available for inter-rater reliability. In the American study, in fact, data collected by two special-education teachers “without a formal language background” [22], and who scored 25 randomly chosen transcripts, were employed to evaluate inter-rater reliability. Their scoring was then compared to the two authors' scores with calculated correlations. The reported inter-

rater reliability scores appear lower than those of the present study. It is possible that these diverging results are related to the number of children involved in the reliability analysis, and to the type of raters. In fact, in the present study a random sample of 178 recordings was listened to and rated by two licensed speech-language pathologists from the same facility, specialized in the assessment and management of language impairment. Another possible cause of the difference in reliability found in the present study compared to the American one lies in the language used. The mean complexity scores in the original and the present study per se do not seem to present a large difference, so we might speculate that reliability in Complexity ratings varies across languages, because of their different grammatical structures. However, to the best of our knowledge, no data exist on this inter-cultural difference in Complexity reliability scoring.

As far as normative data are concerned, I-BST scores obtained in the group of normal developing Italian children appear similar to those previously reported. The standardisation of BST's original version was obtained in a group of 573 children, ages 42–101 months, mostly living in South-East England. For the Information-level measure, in fact, original normative data ranged from 12.61 ± 6.58, for the age class 3;9 years, to 37.20 ± 5.92, for the age class 8;5 years. In our sample, scores ranged from 10.50 ± 4.30, for age class 3; 6-3; 11 years, to 30.40 ± 6.47, for age class 7; 6-7; 11 years.

In line with previous reports [22,23], our sample also maintained a constant, increasing trend for Information, Sentence Length, and Complexity-level scores.

Finally, the analysis of SES showed that the majority of the enrolled families were composed by parents with mixed levels of education- and work-environment. Some differences were found in the mother/father comparison, as fathers were more likely to be manual workers, while mothers were more likely to be office workers. Education level was also observed to be slightly higher in mothers. These differences should not be considered as a bias in the sample, but rather they are representative of typically Italian SES differences. In fact, relevant information available on an Italian national scale (ISTAT, Istituto Nazionale di Statistica) supported the findings of the present study with the majority of men employed as manual workers and the majority of women employed as office workers.

The main limitation of this study is the fact that all the enrolled children were from classrooms in only three different cities in Northern Italy. Greater geographic diversity would have been preferable. In addition, only native-born Italian children were enrolled in the current study, and no information was collected about Italian-speaking children of different ethnicities. Therefore, it is possible that the socioeconomic status of immigrant families and bilingualism both might play a role in the I-BST outcome as previously demonstrated [34].

In reliability analysis, inter-rater and intra-rater reliability included only two assessments for each child, so reliability scores should be considered with caution. Moreover, no measure of concurrent validity was performed, as no other measure of narration was available in Italian. Future studies are needed to further analyze concurrent validity. Another weak point of the study lies in the fact that a rigorous assessment of typical language development was not performed, since parents and teachers of the enrolled children were asked to determine it. Therefore although the normative data can be used in clinical practice, caution should be applied. Finally, the reader is invited to reflect that the normative sample did not reach a number of 100 children for age class; although previous studies presented sample recruited children similar to ours, larger number are required for more rigorous normative data.

5. Conclusions

In conclusion, the current findings support I-BST's reliability and validity for the evaluation of narrative abilities in Italian-speaking children. Normative data from a large cohort of typically developing children may be useful during the clinical evaluation of narrative abilities in both normal and pathological children. I-BST's application is recommended in clinical practice (as part of a battery of instruments and procedures providing information about language development) as well as in epidemiological, efficacy, and outcome research.

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Acknowledgement

the authors are particularly thankful to Bernardi E, De Angeli S, Gravelloni C, Ciceri F, De Angeli S, De Cillis G, Losi S, Rondi S for their help in the assessment of narrative skills, scoring, and data entry. The authors also thank Anne van Kleeck for critical revision of the paper leading to significant improvement in the structure and content of the paper.

References

[1] J.A. Hudson, L.R. Shapiro, From knowing to telling: the development of children's scripts, stories, and personal narratives, in: A. McCabe, C. Peterson (Eds.), Developing Narrative Structure, Erlbaum, Hillsdale, NJ, 1991, pp. 89–136.
[2] B.Z. Liles, Narrative discourse in children with language disorders and children with normal language: a critical review of the literature, J. Speech Hear. Res. 36 (1993) 868–882.
[3] N.W. Nelson, Childhood Language Disorders in Context: Infancy through Adolescence, second ed., Allyn and Bacon, Boston, MA, 1998.
[4] A. McCabe, P.R. Rollins, Assessment of preschool narrative skills: prerequisite for literacy, Am. J. Speech Lang. Pat, A J. Clin. Pract. 13 (1994) 45–56.
[5] L. Feagans, M.I. Appelbaum, Validation of language substyles in learning disabled children, J. Ed. Psyc 78 (1986) 358–364.
[6] J.N. Kaderavek, E. Sulzby, Narrative production by children with and without specific language impairment: oral narratives and emergent readings, J. Speech Lang. Hear. Res. 43 (2000) 34–49.
[7] T. Griffin, L. Hemphill, L. Camp, D. Wolf, Oral discourse in the preschool years and later literacy skills, First Lang. 24 (2004) 123–147.
[8] J.F. Miller, J. Heilman, A. Nockerts, A. Iglesias, L. Fabiano, D.J. Francis, Oral language and reading in bilingual children, Learn. Disabil. Res. Pract. 21 (2006) 30–43.
[9] E. Reese, S. Suggtate, J. Long, E. Schaughency, Children's oral narrative and reading skills in the first 3 years of reading instruction, Read. Writ. 23 (2009) 627–644.
[10] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders (DSM-5), Arlington, VA, 2013.
[11] C.F. Norbury, D. Bishop, Inferential processing and story recall in children with communication problems: a comparison of specific language impairment, pragmatic, language impairment and high functioning autism, Int. J. Lang. Commun. Disord. 37 (2002) 227–251.
[12] N. Botting, Narrative as a tool for the assessment of linguistic and pragmatic impairments, Child Lang. Teach. Ther. 18 (2002) 1–21.
[13] K. Dodwell, E. Bavin, Children with specific language impairment: an investigation of their narratives and memory, Int. J. Lang. Commun. Disord. 43 (2008) 201–218.
[14] C. Miniscalco, B. Hagberg, B. Kadesjo, M. Westerlund, C. Gillberg, Narrative skills, cognitive profiles and neuropsychiatric disorders in 7 and 8-year-old children with late developing language, Int. J. Lang. Commun. Disord. 42 (2007) 665–681.
[15] P. Holck, A. Dahlgren Sandberg, U. Nettelbladt, Narrative ability in children with cerebral palsy, Res. Dev. Disabil. 32 (2011) 262–270.
[16] T. Boons, L. De Raeve, M. Langereis, L. Peeraer, J. Wouters, A. van Wieringen, Narrative spoken language skills in severely hearing impaired school-aged children with cochlear implants, Res. Dev. Disabil. 34 (2013) 3833–3846.
[17] M. Dennis, B. Jacennik, M. Barnes, The content of narrative discourse in children and adolectents after early-onset hydrocephalus and in normally developing age peers, Brain Lang. 46 (1994) 129–165.
[18] D.B. Petersen, S.L. Gillam, R.B. Gillam, Emerging procedures in narrative

assessment: the index of narrative complexity, Top. Lang. Disord. 20 (2008) 111–126.
[19] D. Boudreau, Narrative abilities, advances in research and implications for clinical practice, Top. Lang. Disord. 28 (2008) 99–114.
[20] C.E. Renfrew, The Bus Story: a Test of Continuous Speech, North Place, Old Headington, Oxford, 1969.
[21] S.E. Stothard, M.J. Snowling, D.V.M. Bishop, B.B. Chipchase, C.A. Kaplan, Language-impaired preschoolers: a follow-up into adolescence, J. Speech Hear. Res. 41 (1998) 407–418.
[22] C.E. Renfrew, J. Cowley, C. Glasgow, The Renfrew Bus Story Language Screening by Narrative Recall (American Edition), The Centerville School Delaware, 1994.
[23] C.E. Renfrew, Renfrew Bus Story Test Manual: a Test of Narrative Speech, Winslow, Oxford, 1995.
[24] D.V.M. Bishop, A. Edmundson, Language-impaired 4-year-olds: distinguishing transient from persistent impairment, J. Speech Hear. Res. 52 (1987) 156–173.
[25] G. Conti-Ramsden, N. Botting, Characteristics of children attending language units in England: a national study of 7-year-olds, Int. J. Lang. Commun. Disord. 34 (1999) 359–366.
[26] J.E. Dockrell, G. Lindsay, Children with specific speech and language difficulties – the teachers' perspective, Oxf. Rev. Educ. 27 (2001) 369–394.
[27] A. Gallagher, U. Frith, M.J. Snowling, Precursors of literacy delay among children at genetic risk of dyslexia, J. Child Psychol. Psychiatry 41 (2000) 203–213.
[28] L. Girolametto, M. Wiigs, R. Smyth, E. Weitzman, P.S. Pearce, Children with a history of expressive vocabulary delay: outcomes at 5 years of age, Am. J. Speech-Language Pathology 10 (2001) 358–369.
[29] G. Zarmati, P. Cipriani, F. Mozzanica, R. Salvadorini, Abilità narrative in epoca prescolare con una prova di retelling: studio pilota in soggetti italiani, I Care 3 (2012) 80–87.
[30] D.E. Beaton, C. Bombardier, F. Guillemin, M.B. Ferraz, Guidelines for the process of cross-cultural adaptation of self-report measures, Spine 25 (2000) 3186–3191.
[31] T.J. Cole, P.J. Green, Smoothing reference centile curves: the LMS method and penalized likelihood, Statistics Med. 11 (1992) 1305–1319.
[32] D.V. Hayward, G.E. Stewart, L.M. Phillips, S.P. Norris, M.A. Lovell, Language, Phonological Awareness, and Reading Test Directory, Canadian Centre for Research on Literacy, Edmonton, Alberta, 2008, pp. 1–4. Retrieved from, http://www.uofaweb.ualberta.ca/elementaryed/ccrl.cfm.
[33] M.E. Pankratz, E. Plante, D.M. Insalaco, The diagnostic and predictive validity of the Renfrew Bus Story, Lang. Speech, Hear. Serv. Sch. 38 (2007) 390–399.
[34] A. Van Kleeck, A. Lange, A.L. Schwarz, The effect of race and maternal education level on children's retells of the Renfrew Bus Story-North American Edition, J. Speech Lang. Hear. Res. 54 (2011) 1546–1561.