Psychology of Language and Communication 2024, Vol. 28 No. 1

sciendo

DOI: 10.58734/plc-2024-0002

Paula C. Alves and Ana Paula Vale
Department of Education and Psychology, University of Trás-os-Montes and Alto Douro, Portugal

Replicating the comparison between item-based tests and language samples in Portuguese kindergarteners' language assessment

The current study aimed to examine the relationships between item-based tests and a narrative retelling by analyzing concurrent associations, regression coefficients, language progress, and sex differences. Thirty-one kindergarteners with typical language development were assessed at the beginning (M_{age} = 64.77 months) and the end of the school year. The statistically significant associations and predictive relations between the item-based tests and narrative retelling were scarce and occurred mainly at the same language level (words and sentences). Accordingly, language progression was mirrored differently by the two assessment conditions. Sex was not related to language skills. Overall, the results indicate these two kinds of tools assess different language constructs and should be used complementarily. Additionally, two relatively innovative trends emerged: errors produced in the narrative seem to be a good discriminative metric for this age range, and testing language by units (words and sentences) might be appropriate to assess language development.

Key words: language assessment, item-based tests, language samples, narrative, kindergarten

Address for correspondence: Ana Paula Vale
Department of Education and Psychology, University of Trás-os-Montes and Alto Douro, Quinta de Prados, 5000-801, Vila Real, Portugal.
E-mail: pvale@utad.pt

In kindergarten, children are expected to have acquired much of the oral language structures that an experienced speaker uses (Owens, 2016), and to process it in a sophisticated way (Melzi & Schick, 2017). However, between 7% and 9% of children do not follow the expected development trajectory and show language deficits that will negatively impact their academic achievement and their life course if they are not addressed at an early age (Clegg et al., 2005; Cronin et al., 2020; Le et al., 2020). Accordingly, language difficulties are currently considered a public health problem (Law et al., 2018), indicating that it's crucial to be aware of children's language development profiles before they start formal schooling (UK Department for Education [UK DfE], 2020). Therefore, it is becoming increasingly important to address the issues of language assessment and to acknowledge the differences and complementarities of the tools used for this purpose.

Children's Language Assessment

The support for the perspective that language skills at kindergarten are reliable indicators of further success in learning to read and in the more general academic path has been growing (Guiberson & Ferris, 2019; Hjetland et al., 2019). Also, there is evidence that adulthood achievements are significantly associated with early language skills (Cronin et al., 2020; Forrest et al., 2018; Schoon et al., 2010). Thus, the early detection of language difficulties seems to be one of the most effective ways to minimize and/or avoid major damage at later life stages, since it allows one to take precautionary measures and intervene effectively (Zhang et al., 2020). As a result, assessing language during kindergarten seems decisive (Ebert & Pham, 2017). Accordingly, the UK government has recently funded the assessment, and the respective monitoring, of all Reception Year children (UK DfE, 2020; West et al., 2021), a decision that highlights the fact that early language skills are extremely valuable.

However, assessing language skills and development is a complex task. Thus, language assessments should be carried out based on solid principles, which includes the understanding of how measurement tools can influence children's performances.

Item-Based Assessment

Standardized/item-based tests are some of the most frequently used instruments when assessing language skills, and there are several reasons for relying on such tools. Crucially, they allow peer comparison through norms, hence facilitating the identification of language disorders (Ebert & Scott, 2014; Paul et al., 2018). Additionally, each test usually assesses various skills, and it can be used with different age groups (Ebert & Scott, 2014). Nevertheless, despite their attested value and widespread use, some limitations have been noted (Boerma et al., 2016; Gazella & Stockman, 2003; Paradis et al., 2021; Stockman, 1996). It is worth mentioning that they may constitute an unsatisfactory approach to the lin-

guistic requirements children encounter in everyday life (Gagarina et al., 2019; Stockman, 1996). There is evidence that item-based standardized tests typically reflect language skills that are different from those that can be observed in real language situations, which often involve the use of discourse associated with an activity (Ebert & Scott, 2014). Furthermore, some studies have shown that many item-based tests have reduced ecological validity (Roy & Chiat, 2013). For example, a child may not be able to use the expression "because" in a decontextualized task typical of a standardized test, but they may use it correctly when telling a story. Additionally, the performances that are sanctioned as normative in an item-based test context can be distant from the type of exposure and language experience of many children, particularly those who do not belong to the social groups most commonly represented in mainstream research — middle-class Caucasian children (Kuchirko, 2019). Also, while item-based tests allow for a comprehensive assessment of language since they focus on several domains, this assessment usually lacks in-depth information as each of the domains is tested through a small number of items (Ebert & Scott, 2014).

Narrative Retelling Assessments

Given the limitations of standardized/item-based tests, interest in language samples has increased, and it has been consistently found that narrative tasks are the method most likely to elicit the production of complex utterances (Southwood & Russell, 2004). This advantage is possibly due to the fact that narratives compel children to integrate temporal and causal events (among others) with the production of specific meanings. This prompts the use of a more complex language than the one a child usually produces in dialogues regulated by turn-taking (Humphry et al., 2017). Moreover, some researchers consider that narrative retelling assessments have the potential to generate children's "maximum behavior," and not just their "typical" language behavior (Southwood & Russell, 2004).

However, the need to simultaneously use and combine several language domains in a narrative task while also involving memory processes (Westerveld & Gillon, 2010) could also be an overly complex task for some children compared to item-based tests. Even so, several studies have shown that narrative assessments are an appropriate tool to use with children from diverse linguistic backgrounds (Cleave et al., 2010; Gagarina et al., 2019). Narrative tasks focus on "core language" abilities acquired through basic language exchanges essential for daily life (Roy et al., 2014), therefore being common to the greatest diversity of experiences. Indeed, it has been suggested that narrative assessments are one of the most "ecologically valid ways" in which to evaluate language competence (Boerma et al., 2016).

Lever and Senéchal (2011) investigated the influence of the specific task used in narrative production, comparing telling and retelling with parallel materials. Their results showed that all children produced significantly longer narratives, better structured and with more references to mental states, in the retelling task

than in the telling one. The authors reasoned that a retelling task might require less processing effort than a telling task, thus facilitating a more elaborate use of language. Westerveld and Vidler (2016) also observed that the narrative retellings of five-to-seven-year-olds contained syntactically more complex language than dialogues and personal narratives. On the other hand, narrative retelling could be thought of as essentially a rote repetition task, ambiguously involving the production of linguistic structures which may not be truly acquired by the child. Nevertheless, many studies indicate that this is probably not the case, since they have proved that language repetition tasks reflect linguistic ability (Klem et al., 2015; Nag et al., 2018). It is well acknowledged that children find it difficult to repeat structures that they do not know (Devescovi & Caselli, 2007). As the retelling of narratives seems to activate high-level language skills without introducing too heavy cognitive efforts, it has been increasingly used for assessing language skills (Beytollahi et al., 2020; Diehm et al., 2020; Frizelle et al., 2018; Stipdonk et al., 2020; Westerveld & Vidler, 2016; Westerveld et al., 2012).

Comparison Between Standardized/Item-Based Tests and Narrative Tasks

Given the value of item-based tests and narrative tasks in language assessment, it is important to understand whether these two types of tools can be considered equivalent or if, on the contrary, they assess different language processes.

Several studies have compared narrative samples and item-based tests. Although many indicate that there are areas of convergence between them (Bedore et al., 2010; Ebert & Pham, 2017; Ebert & Scott, 2014; Westerveld et al., 2012), they also show, together with others, that those instruments are not always parallel with each other in reflecting language skills (Ebert & Scott, 2014; Ebert & Pham, 2017; Stipdonk et al., 2020; Westerveld et al., 2012).

When analyzing the findings reported by those studies, a pattern emerged. Among the six correlations involving oral language, Ebert and Scott (2014) observed four significant ones between the mean length of utterance in words (MLU-w) and the subtests of the standardized tests; three of them assessed language at the sentence level. In the study by Ebert and Pham (2017), among the five correlations analyzed, mean length of utterance (MLU) only produced two statistically significant ones, with subtests that were based on sentences. Stipdonk et al. (2020) assessed very preterm (<32 weeks) children and full-term children and analyzed the relationships between six item-based subtests and MLU, but only those that focused on skills at the sentence level (three) reached a statistically significant relationship with MLU. Owens and Pavelko (2017) purposely chose three subtests that elicited responses in sentence format to analyze the relationships with MLU, which proved statistically significant. This pattern of results, in which statistically significant associations are more frequent at the sentence level than at the word level, makes sense because MLU involves the production of sentences. Furthermore, it suggests that sentence processing may constitute a level of

assessment that brings together different types of tools and contexts.

It is also important to emphasize that most studies examining relationships between different assessment instruments included children with language difficulties (Bedore et al., 2010; Ebert & Pham, 2017; Ebert & Scott, 2014). This may have influenced the reported associations between tools and emphasizes the idea that there is still more to explore regarding the possible complementarities and/or differences between different types of instruments when assessing children with a typical development path.

Like with item-based tests, previous studies have found associations between age and several narrative measures (Owens & Pavelko, 2020; Pavelko & Owens, 2017; Rice et al., 2010; Voniati, 2016; Voniati et al., 2020; Westerveld & Vidler, 2016; Wu, 2020). However, according to Voniati (2016), although MLU-w may be an index that increases with age, it may also depend on experience. Voniati (2016) puts forward the hypothesis that individual differences in rates of development and the complexity of the language that is being assessed may also account for discrepancies in findings relating age to MLU. Illustratively, Santos et al. (2015) did not find significant differences in MLU between the 4- and 5-year-old groups in European Portuguese. Thus, the relationship between age and narrative measures should be taken with caution, given that information is still scarce for many languages.

In examining sex differences in language development, prior studies have differed in the age ranges, sample sizes and tools used (Rinaldi et al., 2021). Furthermore, the authors also highlight that, despite several studies presenting statistically significant differences between sexes, the usual advantage in girls has often remained small and is not present in all the ages that have been assessed. In effect, this advantage seems to be more evident in early ages and tends to progressively decrease when a given ability is already acquired by most children.

Given the inconsistency of the above-mentioned outcomes, there is still a need to better understand the relationships between the item-based tests and narrative samples. This is markedly so for the European Portuguese language, for which we are aware of only one study using these two types of assessment tools (Santos et al., 2015). The authors assessed 92 children, aged between 4;00 and 5;05 years, who were Portuguese native speakers and had not been included in speech or language therapy. Their results regarding the associations between the item-based test and the narrative assessment led the authors to state that statistically significant correlations are expected only until the end of kindergarten for Portuguese children.

To examine associations between different assessment tools, we tested a group of Portuguese kindergarteners with typical language development, in two moments during the school year (T1 and T2), with the following objectives: (a) to examine to what extent language skills measured through language samples are related to those measured through standardized/item-based tests, (b) to examine whether item-based language scores in T1 might predict narrative retelling skills in T2 and whether narrative retelling in T1 might predict item-based language

scores in T2, (c) to examine the progress of the participants' language levels over a period of about six months using different instruments, and (d) to analyze differences in language skills as a function of sex using different instruments.

Therefore, it was hypothesized that:

1) The associations between the instruments will be stronger between measures that index the same level of language (words, sentences);

2) In measures that index the same level of language, item-based test scores in T1 will predict narrative retelling in T2 and narrative retelling in T1 will predict item-based test scores in T2;

3) There will be statistically significant progress in children's language in the item-based tests but not in the measures extracted from the narrative retelling;

4) There will not be a difference in language skills between sexes across assessment tools.

Method

Participants

The vast majority of Portuguese children enroll in preschool at three years of age and in kindergarten between the ages of five and six, entering primary school between the ages of six and seven in 1st grade. As is typical for other languages, at the end of kindergarten children are expected to master a diverse vocabulary and use complex sentences, goals that are inscribed in the government guidelines for kindergarten (Silva et al., 2016).

Participants attended three kindergarten public school classes in a town located in the inner north of Portugal. Only monolingual children who presented a typical language development were included — that is, children who reached scores at or above the standardized test's (TL-ALPE; Mendes et al., 2014) cutoff point. Participants came from a low-to-middle socio-economic status and were not enrolled in any type of language therapy. We did not control for how long each child had been involved in the educational system but, typically, the vast majority of children enroll in public school (kindergarten) by the age of three. However, since mandatory schooling in Portugal only starts in the 1st grade, we cannot reject the possibility that the schooling experience may have been different to some extent among the participants.

Initially, there were 33 children, however, one child did not complete all the tests in the T1, so this child was excluded from the study. Later, another participant was eliminated due to outlier scores. Given this child's behavior during the assessment process, it was considered that their scores did not reflect their true skills and, consequently, they were not reliable. Thus, the final sample (see Table 1) included 31 children, 19 boys, with a mean age of 64.77 months in T1 and a mean age of 70.29 months in T2.

Table 1. *Sociodemographic Characteristics of the Participants*

	Boys	Girls	*M* (*SD*)	Min	Max
Sex	19 (61.3%)	12 (38.7)	-	-	-
Age in T1	-	-	64.77 (2.513)	59	68
Age in T2	-	-	70.29 (2.506)	63	72

Note. M = mean; *SD* = standard deviation; Min = minimum; Max = maximum

Instruments and Procedure

After obtaining all the ethically necessary authorizations, data collection was carried out in a quiet room in the children's schools and occurred in two time points. T1 occurred at the beginning of the school year (between October and November) and T2 at the end of the school year (between April and May). The children were assessed individually using the same four instruments during both time points.

The current study replicated the examination of the relationships between the scores obtained in a standardized item-based test and narrative retell measures for the European Portuguese language. In addition to that comparison, expressive and receptive vocabulary item-based assessment instruments were also used. Unlike the standardized test, which gathers different units of language – words and sentences – into a single result, vocabulary tests exclusively assess the processing of words. This difference between the standardized test and the vocabulary tests will allow us to examine the different types of relationships with the measures to be extracted from the narrative, which also focus on words and sentences but uttered in the context of discourse production.

Language Test – Pre-Primary School Language Assessment (Teste de Linguagem – Avaliação de linguagem pré-escolar, TL-ALPE)

This is a standardized test (Mendes et al., 2014) that aims to assess the Portuguese language skills of children between the ages of three and six. It includes receptive (CA) and expressive oral language (EVO) sections, which assess the semantic and morphosyntactic domains. As usual, the majority of the EVO subtests (nine out of 13) allow a word or phrase as an answer. Of the 16 subtests (receptive and expressive), only four elicit the production of sentences, and only a total of seven items (three in EVO and four in CA) involve the production of complex sentences. In a validation study, the TL-ALPE obtained an $\alpha = .83$ for receptive language and an $\alpha = .94$ for expressive language, which means that it presents good to very good internal consistency (Marôco & Garcia-Marques, 2006). For scoring, each correct answer was given one point.

Peabody Picture Vocabulary Test (PPVT-4)

Participants completed an adapted and reduced version (Vicente et al., 2011a; 2011b) of the original PPVT (Dunn & Dunn, 2007), which is a receptive vocabulary test. In this test, the children heard, one by one, 35 words, previously recorded by a native speaker of European Portuguese. For each of the words, four images were displayed on the screen of a computer, and the children were asked to choose the one that best matched the word they had heard. Correct answers were quoted with a point. In the current study, the PPVT-4 obtained an $\alpha = .64$ at T1 and an $\alpha = .66$ at T2, meaning it presented a satisfactory internal consistency (Marôco & Garcia-Marques, 2006).

Expressive Vocabulary Test (Teste de vocabulário expressivo, TVE)

This is an experimental task where children were asked to provide the names of the objects, animals, buildings, and actions in pictures presented on a computer screen. Children completed two training items before starting the test. The answers were recorded in audio format and were scored later. The test contained 40 images, and one point was given for each correct answer. This task differs from ALPE-EVO in that it requires children to answer using only a single word, whereas EVO combines answers produced with single words and with sentences, considering the assessment of vocabulary, category naming, and use of antonyms. Regarding internal consistency, in the current study, the TVE obtained an $\alpha = .76$ in T1 and an $\alpha = .78$ in T2, which are considered good psychometric values (Marôco & Garcia-Marques, 2006).

Narrative Retelling in Portuguese (Reconto da narrativa em Português, RN-P)

Four colored images (see the Appendix for Image 1) were presented in a chronologically ordered sequence to each child, accompanied by the reading of a story. The story was about a boy that is playing tag in the schoolyard and has one shoe untied. One of his friends steps on his shoe and it falls off his foot. Then, it starts to rain, and the boy feels worried because his sock is wet, and he does not know where his shoe fell. He asks his teacher for help and they both find the shoe, putting it back on. Lastly, his teacher advises him to always tie his shoes. Immediately after, each child was asked to retell the story as they looked at the images. Throughout the retelling, non-specific prompts could be provided for each image, such as "hm, hm" or "and then?", but not direct questions.

Narrative Measures

Mean length of utterance is one of the most used measures in the analysis of narratives, for clinical and research purposes (Tomas & Dorofeeva, 2019). Ad-

ditionally, MLU is a relevant index of morphosyntactic complexity (Pavelko & Owens, 2017) and it is statistically significantly associated with vocabulary diversity (DeThorne et al., 2005). Therefore, MLU is considered one of the most accurate measures to assess the level of whole language development (Brown, 1973; Rice et al., 2010).

The most common ways to calculate MLU are the counting of morphemes (MLU-m) or words (MLU-w) per utterance (Ezeizabarrena & Fernandez, 2018). The identification of morphemes and, consequently, their counting, can be a challenging and inconclusive task, particularly when dealing with zero and/or bound morphemes, which generate frequent analysis disagreements in languages with moderate to high morphological complexity, as is the case of the Portuguese language. Concurrently, several studies have reinforced the relevance of using MLU--w. Illustratively, Bishop (2004) found a steady rise in MLU-w as age increased, and diverse experiments have used the MLU-w from the moment children were able to produce complete sentences (Frizelle et al., 2018). In addition, high correlations have been found between MLU-m and MLU-w in languages with varying degrees of morphological complexity, such as French (Mimeau et al., 2015), Spanish (Aguado, 1995), Basque (Ezeizabarrena & Fernandez, 2018), Dutch, Icelandic, and Irish (Parker & Brorson, 2005). Thus, since MLU-w is calculated more easily and significantly reduces the subjectivity of the decisions to be made, analyzing MLU in words may be preferable (Ezeizabarrena & Fernandez, 2018).

Another language index that can be collected through narrative tasks is lexical diversity. Lexical diversity refers to the variety of vocabulary items a speaker uses, which indicates their developmental level. It can be measured in several ways. Among the alternatives – number of different words (NDW), diversity of vocabulary (VocD), moving-average type-token ratio (MATTR) – MATTR has proven to be the most valid and reliable measure since it is reputed to be unbiased by the sample size (Luckman et al., 2020; MacWhinney & Fromm, 2016). The procedure to calculate MATTR involves the use of a movable window through the speech sample, which estimates the TTRs (type-token ratios, that is, the ratio of the number of different words [types] by the total number of words [tokens]) for each successive window with a fixed length (Luckman et al., 2020). For example, if we choose a window with a length of 10 words, then the program will calculate the TTR of the words 1-10, then 2-11, then 3-12, and so on, until the end of the speech sample. Afterwards, the average of all these TTRs is calculated. Thus, a MATTR closer to 1.0 indicates a varied vocabulary, and a MATTR closer to 0.0 represents a limited, repetitive vocabulary.

In the current study, the narrative retellings were recorded in audio format and were later transcribed into a text file. Then, the narratives were coded using the CHAT system (MacWhinney, 2000) and were processed by the CLAN program (MacWhinney, 2000), which produced data for MLU-w, MATTR, and the total number of words (TNW) of all the children in each of the two moments of the assessment. A manual analysis of each child's narrative was also performed

for each phase of assessment, in which the total of word errors and non-sentences (TENS) and the total of subordinated clauses (TSC) were counted and used as measures of grammatical complexity. Other studies have also used measures that involved grammatical errors (e.g. Ebert & Pham, 2017; Ebert & Scott, 2014) and subordinated clauses (e.g. Frizelle et al., 2018)

Considering previous studies (e.g., Pavelko & Owens, 2017; Santos et al., 2015; Voniati, 2016; Westerveld et al., 2012), the segmentation of the utterances was based on syntax and prosodic clues, namely, pauses and intonation changes. Thus, a new utterance was counted when there was a change of intonation, a long pause (at least two seconds) between two statements, a change of subject in the next statement, and the use of the expression "and then". The production of "and" between statements was considered as a marker of a new utterance only if the subject in the second clause was explicitly expressed. Sentences were also separated into two (or more) short utterances, even when "and" or other conjunctions were not used. For example, "(Ele) perdeu o sapato, / (ele) não via o sapato" ([He] lost his shoe, / [he] didn't see the shoe"). Change of utterance is marked by /. "He" (Ele) is not needed in Portuguese for the utterance to be grammatically correct as the subject is implied. Additionally, subordinate clauses (linked by "because," "but," "then," "who," or "when," among others) were maintained in a single utterance if the ideas were related. In European Portuguese, a subordinate clause can act on the main clause as a noun, an adjective or an adverb clause.

Two encoders separately analyzed 80% of the transcriptions and encodings of the RN-P and, after that, there were discussion exchanges for each language sample between the encoders until agreement was achieved.

Data Analysis

The instruments' internal consistency analyses were carried out using Cronbach's α values (Marôco & Garcia-Marques, 2006).

To examine the assumptions of normality, the coefficients of skewness and kurtosis were analyzed (George & Mallery, 2016). Three variables did not follow a normal distribution, namely, TSC in T1, TL-ALPE-CA, and TL-ALPE-total in T2, so nonparametric statistics were used to analyze them.

To correct the effects of multiple testing, a Bonferroni correction was applied to the analyses when needed.

Results

Descriptive Analysis

As shown in Table 2, children had high but typical levels of standardized scores on language assessed by the TL-ALPE. Table 2 also shows the raw scores in the TL-ALPE-EVO, TL-ALPE-CA and TL-ALPE-total (henceforth, ALPE-EVO,

Table 2. *Descriptive Values for Each Variable and Comparison of the Language Levels Between the Two Assessment Time Points with a Bonferroni Correction*

	Assessment moment	*M*±*SD*	Skw.	Kur.	Min.	Max.	*t*/*W* (*df*)	*p*	*d*/rC
ALPE-EVO points	T1	58.61±4.794	−0.525	−0.627	49	66	−7.301 (30)	**.000***	1.465
	T2	65.19±5.062	−1.214	1.487	50	73			
ALPE-EVO standardized	T1	102.74±9.784	−0.391	−0.111	79	121	−4.990 (30)	.000	−0.896
	T2	112.65±10.729	−1.035	1.065	84	128			
ALPE-CA points	T1	44.42±2.814	−0.470	−0.228	38	49	3.860	**.000***	0.903
	T2	47.29±1.883	−1.699	2.519	42	49			
ALPE-CA standardized	T1	105.65±13.258	-0.327	-0.367	78	129	3.612	.000	−0.781
	T2	117.77±8.966	−1.605	2.224	93	127			
ALPE-total points	T1	103.03±6.843	−.244	−.856	88	114	−4.658	**.000***	1.432
	T2	112.48±6.470	−1.370	2.132	92	122			
PPVT-4	T1	27.42±3.413	−0.293	-0.325	20	34	−5.818 (30)	**.000***	1.311
	T2	30.97±2.811	−0.514	-0.857	25	35			
TVE	T1	26.90±5.394	−0.936	0.879	12	37	−8.155 (30)	**.000***	1.045
	T2	30.55±5.208	−0.955	0.586	17	38			
MLU-w	T1	5.949±1.182	0.232	−0.317	3.714	8.700	−0.572 (30)	.572	−0.103
	T2	6.096±1.005	0.258	−0.212	4.333	8.333			
MATTR	T1	.813±0.058	-0.031	−0.156	.698	.943	0.070 (30)	.944	0.013
	T2	.812±0.060	-0.005	−0.002	.686	.953			
TENS	T1	0.936±0.814	0.519	−0.213	0	3	1.278 (30)	.211	0.229
	T2	0.677±0.645	1.477	1.483	0	3			
TSC	T1	1.160±1.369	2.692	10.548	0	7	−0.017 (30)	.987	0.059
	T2	1.060±1.031	0.645	−.652	0	3			
TNW	T1	48.581±13.691	0.938	0.662	26	87	−0.759 (30)	.454	−0.136
	T2	51.065±15.860	0.403	−0.234	22	87			

Note. * significant at $p < .004$, according to a Bonferroni correction. *M* = Mean; *SD* = Standard-deviation; Skw = Skewness; Kur = Kurtosis; Min = Minimum; Max = Maximum; *t* = *t*-student's test statistic; *W* = Wilcoxon's test statistic; *df* = degrees of freedom; *p* = probability of significance; *d* = Cohen's d; rC = matched-pairs rank-biserial correlation; ALPE-EVO = TL-ALPE's oral verbal expression component; ALPE-CA = TL-ALPE's receptive understanding component; ALPE-total = TL-ALPE's total score; PPVT-4 = Peabody Picture Vocabulary Test - 4; TVE = Expressive Vocabulary task; MLU-w = Mean length of utterance – words; MATTR = Moving-average type-token ratio; TENS = Total of Word Errors and Non-Sentences; TSC = Total of Subordinated Clauses; TNW = Total number of words.

ALPE-CA, and ALPE-total), since these were the results used in the analyses. MLU-w and the TNW in T1 were very similar to those obtained by Sapage et al. (2019), also using a four-image narrative retelling task with Portuguese five-y-ear-olds (5.56 and 46.00, respectively), suggesting that the present results may be typical for Portuguese children of these ages.

Concurrent Correlations

As shown in Tables 3 and 4, the ALPE scores produced significant correlations of moderate magnitude with one single measure, the TENS extracted from the narrative samples, only in T1. All the other correlations between the RN-P measures and the results of the item-based tests did not reach statistical significance, either in T1 or T2.

Comparison of Language Levels Between the Two Moments of Assessment

Table 2 shows that language levels improved from T1 to T2 when assessed by the TL-ALPE, PPVT-4 and TVE, since they presented statistically significant differences between the two assessment time points, with strong effects.

On the contrary, for the RN-P measures, there were no statistically significant differences between T1 and T2, meaning that language skills produced in the narrative retelling context did not progress noticeably.

Table 3. *Concurrent Correlations Between the Item-Based Tests Scores and Measures Obtained from the Narrative in T1*

	MLU-w [CI]	MATTR [CI]	TENS [CI]	TSC [CI]	TNW [CI]
ALPE-EVO	.243 [−.122, .550]	.005 [−.390, .360]	**−.374*** [−.698, .011]	−.146 [−.473, .213]	.146 [−.210, .446]
ALPE-CA	.086 [−.316, .420]	.183 [−.263, .553]	**−.497**** [−.715, -.313]	−.119 [−.448, .233]	−.052 [−.411, .258]
ALPE-total	.206 [−.182, .527]	.079 [−.330, .445]	**−.466**** [−.716, -.173]	−.128 [−.446, .192]	.081 [−.284, .404]
PPVT-4	.138 [−.161, .467]	.312 [.040, .588]	-.062 [−.419, .355]	−.059 [−.386, .275]	.007 [−.410, .342]
TVE	.240 [−.051, .527]	.090 [−.272, .404]	.029 [−.443, .375]	−.020 [−.362, .371]	−.005 [−.318, .248]

Note. *significant at $p < .05$; **significant at $p < .01$. CI = Confidence interval; ALPE-EVO = TL-ALPE's oral verbal expression component; ALPE-CA = TL-ALPE's receptive understanding component; ALPE-total = TL-ALPE's total score; PPVT-4 = Peabody Picture Vocabulary Test - 4; TVE = Expressive Vocabulary task; MLU-w = Mean length of utterance – words; MATTR = Moving-average type-token ratio; TENS = Total of Word Errors and Non-Sentences; TSC = Total of Subordinated Clauses; TNW = Total number of words.

Table 4. *Concurrent Correlations Between the Item-Based Tests Scores and Measures Obtained from the Narrative in T2*

	MLU-w [CI]	MATTR [CI]	TENS [CI]	TSC [CI]	TNW [CI]
ALPE-EVO	−.257 [−.583, .228]	.011 [−.354, .371]	**−.049 [−.348, .273]**	.010 [−.310, .326]	−.030 [−.394, .426]
ALPE-CA	−.159 [−.521, .226]	.127 [−.223, .461]	**−.098 [−.435, .235]**	−.061 [−.363, .292]	−.197 [−.483, .108]
ALPE-total	−.146 [−.478, .228]	−.113 [−.462, .250]	**−.012 [−.329, .325]**	−.051 [−.394, .316]	.002 [−.336, .336]
PPVT-4	−.009 [−.441, .353]	−.072 [−.468, .390]	.147 [−.121, .376]	.116 [−.186, .412]	−.145 [−.482, .199]
TVE	−.062 [−.445, .438]	−.110 [−.492, .287]	.139 [−.126, .408]	.049 [−.247, .351]	.180 [−.230, .621]

Note. CI = Confidence interval; ALPE-EVO = TL-ALPE's oral verbal expression component; ALPE-CA = TL-ALPE's receptive understanding component; ALPE-total = TL-ALPE's total score; PPVT-4 = Peabody Picture Vocabulary Test - 4; TVE = Expressive Vocabulary task; MLU-w = Mean length of utterance – words; MATTR = Moving-average type-token ratio; TENS = Total of Word Errors and Non-Sentences; TSC = Total of Subordinated Clauses; TNW = Total number of words.

Table 5. *The Predictive Role of T1's Item-Based Scores on Narrative Retelling Measures in T2*

		T2 MLU-w	T2 MATTR	T2 TENS	T2 TSC	T2 TNW
Model	Adjusted R^2	.038	**−.009**	.305	.013	−.108
	F	1.299	**0.935**	4.286	1.096	0.267
	p	.296	**.459**	**.009***	.379	.896
T1 ALPE-EVO	B [CI]	−0.114 [−0.226, −0.003]	0.002 [−0.005, 0.009]	−0.002 [−0.091, 0.087]	−0.080 [−0.196, 0.036]	0.037 [−1.853, 1.927]
	SE	0.054	0.003	0.043	0.056	0.919
	β	−.546	.162	−.009	−.372	.011
	t	−2.108	0.611	−0.039	−1.417	0.040
	p	.045*	.547	.969	.168	.969
T1 ALPE-CA	B [CI]	0.085 [−0.081, 0.251]	0.003 [−0.007, 0.013]	−0.154 [−0.287, −0.021]	0.099 [−0.074, 0.271]	0.685 [−2.128, 3.498]
	SE	0.081	0.005	0.065	0.084	1.369
	β	.239	.129	−.458	.270	.122
	t	1.055	0.558	−2.382	1.177	0.500
	p	.301	.582	**.025***	.250	.621
T1 PPVT-4	B [CI]	−0.056 [−0.180, 0.068]	0.003 [−0.005, 0.011]	0.153 [0.054, 0.253]	0.054 [−0.075, 0.183]	−0.887 [−2.991, 1.216]
	SE	0.060	0.004	0.048	0.063	1.023
	β	−.190	.167	.554	.179	−.191
	t	−0.928	0.795	3.177	0.861	−0.867
	p	.362	.434	**.004****	.397	.394
T1 TVE	B [CI]	0.048 [−0.043, 0.139]	−0.005 [−0.010, 0.001]	−0.036 [−0.109, 0.036]	0.046 [−0.048, 0.140]	0.164 [−1.372, 1.700]
	SE	0.044	0.003	0.035	0.046	0.747
	β	.258	−.409	−.206	.240	.056
	t	1.090	−1.687	−1.022	1	0.219
	p	.286	.103	.316	.326	.828

Note. *significant at $p < .05$; **significant at $p < .01$. MLU-w = Mean length of utterance – words; MATTR = Moving-average type-token ratio; TENS = Total of Word Errors and Non-Sentences; TSC = Total of Subordinated Clauses; TNW = Total number of words; ALPE-EVO = TL-ALPE's oral verbal expression component; ALPE-CA = TL-ALPE's receptive understanding component; PPVT-4 = Peabody Picture Vocabulary Test - 4; TVE = Expressive Vocabulary task; Adjusted R^2 = adjusted coefficient of determination; *F* = ANOVA's test statistic; *p* = probability of significance; B = unstandardized beta; CI = Confidence interval; SE = Standard deviation; β = standardized beta; *t* = *t* student's test statistic.

Predictive Analyses

To analyze whether item-based language scores in T1 might predict narrative retelling in T2, a multiple regression analysis was carried out for each of the narrative measures as a dependent variable. The model included T1's ALPE-EVO, ALPE-CA, PPVT-4 and TVE as predictor variables. T1's ALPE-total was excluded given its high correlations with ALPE-EVO and ALPE-CA (r > .80), sugge-

Table 6. *The Predictive Role of T1's Narrative Retelling Measures on Item-Based Scores Measures in T2*

		T2 ALPE-EVO	T2 ALPE-CA	T2 PPVT-4	T2 TVE
Model	Adjusted R^2	−.126	−.126	−.173	.083
	F	0.329	0.330	0.116	1.545
	p	.890	.890	.988	.212
T1 MLU-w	B [CI]	1.139 [−0.950, 3.229]	−0.255 [−1.032, 0.522]	0.148 [−1.036, 1.332]	1.987 [0.047, 3.926]
	SE	1.015	0.377	0.575	0.942
	β	.266	−.160	.062	.451
	t	1.123	−0.676	0.257	2.109
	p	.272	.505	.799	**.045***
T1 MATTR	B [CI]	10.371 [−29.692, 50.434]	3.774 [−11.127, 18.676]	7.470 [−15.231, 30.171]	16.484 [−20.705, 53.673]
	SE	19.452	7.235	11.023	18.057
	β	.118	.115	.153	.182
	t	0.533	0.522	0.678	0.913
	p	.599	.606	.504	.370
T1 TENS	B [CI]	−0.060 [−2.632, 2.511]	−0.401 [−1.358, 0.555]	−0.070 [−1.527, 1.387]	0.608 [−1.779, 2.995]
	SE	1.248	0.464	0.707	1.159
	β	−.010	−.173	−.020	.095
	t	−0.048	−0.864	−0.099	0.525
	p	.962	.396	.922	.604
T1 TSC	B [CI]	−0.054 [−2.030, 1.923]	0.031 [0.704, 0.767]	−0.225 [−1.344, 0.895]	−0.673 [−2.507, 1.162]
	SE	0.960	0.357	0.544	0.891
	β	−.014	.023	−.109	−.177
	t	−0.056	0.088	−0.413	−0.755
	p	.956	.930	.683	.457
T1 TNW	B [CI]	−0.050 [−0.247, 0.147]	−0.001 [−0.074, 0.073]	0.018 [−0.094, 0.129]	0.072 [−0.111, 0.255]
	SE	0.096	0.036	0.054	0.089
	β	−.135	−.005	.086	.190
	t	−0.521	−0.018	0.327	0.812
	p	.607	.985	.746	.424

Note. *significant at $p < .05$; **significant at $p < .01$. MLU-w = Mean length of utterance – words; MATTR = Moving-average type-token ratio; TENS = Total of Word Errors and Non-Sentences; TSC = Total of Subordinated Clauses; TNW = Total number of words; ALPE-EVO = TL-ALPE's oral verbal expression component; ALPE-CA = TL-ALPE's receptive understanding component; PPVT-4 = Peabody Picture Vocabulary Test - 4; TVE = Expressive Vocabulary task; Adjusted R^2 = adjusted coefficient of determination; *F* = ANOVA's test statistic; *p* = probability of significance; B = unstandardized beta; CI = Confidence interval; SE = Standard deviation; β = standardized beta; *t* = *t* student's test statistic.

sting multicollinearity.

As shown in Table 5, the only statistically significant model was the one concerning the prediction of T2 TENS scores, with a moderate predictive value. In this model, there were only two significant predictors. For each point earned in

ALPE-CA in T1, the TENS diminished by 0,15 points in T2. For PPVT-4 in T1, each point explained an increase of 0,15 points in T2's TENS. Also, the MLU-w in T2 was statistically significantly predicted by ALPE-EVO in T1. However, this single result should be interpreted with caution since it occurred in a non-significant model, which is a more powerful statistic.

To analyze whether narrative retelling at T1 might predict item-based scores at T2, a multiple regression analysis was carried out for each of the item-based measures as a dependent variable. The model included T1's MLU-w, MATTR, TENS, TSC and TNW as predictor variables. As can be observed in Table 6, none of the models were statistically significant which means that none of the item-based tests scores at T2 was explained by the measures obtained from the narratives at the beginning of kindergarten. T2's TVE was statistically significantly predicted by T1's MLU-w but, since this result occurred in a non-significant model, it should also be interpreted with caution.

Differences in Language Skills by Sex

Table 7 shows that considering a Bonferroni correction ($p = .0025$), no statistically significant differences were found between sexes in the language measures in either of the two assessment moments. The Hedges' g and rank-biserial correlation values revealed mostly weak effect magnitudes, except at T2's ALPE-CA, which presented a high magnitude.

Discussion

The main objective of the current study was to examine the relationships between measures extracted from a narrative retelling and those obtained through item-based tests, such as the TL-ALPE, which is a standardized generalist test, and the PPVT-4 and TVE, two vocabulary assessment tools.

Our main results show that for Portuguese children between the ages of 4;11 to 6;0 with good language levels, there were only a few statistically significant relationships between language scores resulting from item-based assessments and the measures extracted from language samples produced in narrative retelling.

The statistically significant correlations occurred at the beginning of kindergarten between the item based ALPE scores (all three of them) and the total number of errors and non-sentences (TENS), a measure of the grammatical complexity obtained on the narrative samples and the correctness of word production, which is independent of the narrative content. No statistically significant relations were found for the assessments run at the end of the school year. This partially validates, only for the beginning of kindergarten, our first hypothesis, as TENS is the only narrative metric that combines words (errors) and sentences (non-sentences) as ALPE does. Therefore, the only statistically significant relations observed between the two types of tools tapped the same language units (words and sentences).

Table 7. *Differences in the Measures Extracted from the Different Instruments in Function of Sex in T1 and T2 with a Bonferroni Correction*

	Sex	*M*±*SD*	*t*/*U*(*df*)	*p*	g/rg
ALPE-EVO (T1)	Female	59.92±4.188	1.213 (29)	.235	0.436
	Male	57.79±5.073			
ALPE-EVO (T2)	Female	66.42±3.988	1.072 (29)	.293	0.385
	Male	64.42±5.601			
ALPE-CA (T1)	Female	44.58±2.968	.254 (29)	.801	0.091
	Male	44.32±2.790			
ALPE-CA (T2)	Female	48.25±0.754	−2.380	.023	**0.874**
	Male	46.68±2.136			
ALPE-total (T1)	Female	104.50±6.274	.947 (29)	.351	0.340
	Male	102.11±7.187			
ALPE-total (T2)	Female	114.67±4.313	−1.262	.220	0.548
	Male	111.11±7.295			
PPVT-4 (T1)	Female	28.00±2.412	.831 (28.995)	.413	0.268
	Male	27.05±3.937			
PPVT-4 (T2)	Female	31.75±2.734	1.243 (29)	.224	0.446
	Male	30.47±2.816			
TVE (T1)	Female	28.17±5.540	1.038 (29)	.308	0.373
	Male	26.11±5.290			
TVE (T2)	Female	31.00±6.179	.378 (29)	.708	0.136
	Male	30.26±4.653			
MLU-w (T1)	Female	5.965±1.362	0.061 (29)	.952	0.022
	Male	5.938±1.092			
MLU-w (T2)	Female	5.993±1,089	−0.445 (29)	.659	0.160
	Male	6,161±0.974			
MATTR (T1)	Female	.808±0.067	−0.647 (29)	.731	0.125
	Male	.816±0.053			
MATTR (T2)	Female	.804±0.062	−0.557 (29)	.582	0.200
	Male	.817±0.060			
TENS (T1)	Female	0.917±0.793	−0.101 (29)	.921	0.036
	Male	0.947±0.848			
TENS (T2)	Female	0.500±0.674	−0.827 (29)	.415	0.297
	Male	0.790±1.084			
TSC (T1)	Female	1.420±1.929	−0.214	.857	0.295
	Male	1.000±0.882			
TSC (T2)	Female	0.750±0.866	−1.370 (29)	.181	0.492
	Male	1.260±1.098			
TNW (T1)	Female	79.170±16.672	0.186 (29)	.853	0.067
	Male	48.210±11.924			
TNW (T2)	Female	47.500±12.866	−0.994 (29)	.328	-0.357
	Male	53.320±17.442			

Note. t = t student's test statistic; U = Mann-Whitney's test statistic; df = degrees of freedom; p = probability of significance; g = Hedges' g; rg = rank-biserial correlation; ALPE-EVO = TL-ALPE's oral verbal expression component; ALPE-CA = TL-ALPE's receptive understanding component; ALPE-total = TL-ALPE's total score; PPVT-4 = Peabody Picture Vocabulary Test - 4; TVE = Expressive Vocabulary task; MLU-w = Mean length of utterance – words; MATTR = Moving-average type-token ratio; TENS = Total of Word Errors and Non-Sentences; TSC = Total of Subordinated Clauses; TNW = Total number of words.

The idea that the language unit (sentences or words) being processed may play a role in associating language skills obtained from different assessment tools helps to explain, at least in part, our findings. Indeed, of the 16 TL-ALPE subtests (receptive and expressive), only four require the production of sentences, and only a total of seven items involve the production of complex sentences. These features of the test, which prompt answers of only one word in 12 subtests, have possibly limited the associations with MLU-w results. In support of this perspective, recent studies have indicated that processing at the level of "combinations" and the lexical level are interdependent skills but constitute two different language dimensions (Lonigan & Milburn, 2017; Mouzaki et al., 2020). The same rationale applies to the absence of other statistically significant relationships between the narrative metrics and the item-based test scores.

Following a complementary line of reasoning, many studies (Ebert & Pham, 2017; Ebert & Scott, 2014; Norbury & Bishop, 2003; Westerveld et al., 2012) also support the idea that as age and language development progress, the relations between measures extracted from narrative retellings and item-based test scores become less significant. Our results concur with these findings. At T2, when the children were about 6;0 years old and showed high language skills in the standardized test, the correlations found for MLU-w and for the TNW of the narratives were not statistically significant and were negative. This outcome is in line with the results from Stipdonk et al. (2020) showing that children with higher language scores in standardized tests tend to use more complex and, therefore, shorter sentences in narratives. The perspective that the strength of associations between language measures obtained with different types of instruments weakens with language progress aligns with evidence that language has probably a one--dimensional nature at earlier ages, evolving to a multidimensional one from the end of kindergarten, a time point in which discourse skills emerge as a singular dimension (Karlsen et al., 2021; Lonigan & Milburn, 2017; Mouzaki et al., 2020). Thus, our findings indicate that the use of words in sentences reflects a language knowledge different from that used in tasks eliciting the use of isolated words in formal, closed contexts as those provided by item-based tests.

Regarding the lexical diversity measured by MATTR, an index of the sophistication of the vocabulary used in the narrative, we also did not obtain statistically significant relations with the standardized test, nor with the item-based vocabulary tests. Most studies analyzed the relationships between vocabulary tests and NDW (Bedore et al., 2010; Ebert & Pham, 2017; Ebert & Scott, 2014; Luckman et al. 2020; Ukrainetz & Blomquist, 2002; Westerveld et al., 2012), and found a consistent pattern indicating mainly statistically non-significant associations. Narrative retelling tasks require children to use vocabulary that describes and connects events in a relatively circumscribed way to the model to which they were exposed. This circumstance may limit the relationships with tasks that elicit diversity in isolated lexical production, such as item-based tests do. This kind of relationship may be particularly distant when children have a typical develop-

ment, as was the case with the children in our sample. These children can recall and recount the narratives with great fidelity, producing a high degree of homogeneity among themselves that reduces the strength of the relationships with tests designed to assess small individual differences.

The second hypothesis was also partially confirmed. We hypothesized that in measures that index the same language unit (words and sentences), item-based tests scores obtained at the beginning of the school year would predict narrative metrics at the end of the school year and, vice versa, narrative measures would predict item-based test scores. Nevertheless, only item-based tests were statistically significant predictors. This is in accordance with recent evidence showing that narrative skills are fragile or statistically non-significant predictors (Karlsen et al., 2021; Westerveld et al., 2012) of language aspects measured by structured item-based tests.

Reinforcing the concurrent associations discussed above, only one narrative measure was statistically significantly predicted by structured tests, the TENS, a hybrid metric combining the production of non-sentences and word errors, a measure that was independent of the narrative size and content. The predictor, the ALPE-CA, an item-based subtest which combines listening comprehension of words and sentences, as well as the production of sentences required to answer questions, explained individual differences displayed on the TENS at the end of kindergarten. Indeed, there is evidence that receptive language may be a stronger correlate and predictor of expressive language than measures of expressive language in kindergarten children (Chiat & Roy, 2008; Yarian et al., 2021).

Receptive vocabulary, as measured by the PPVT-4, also explained individual differences on the same narrative metric, the TENS. Systematic relations between receptive vocabulary and grammar and word production at kindergarten ages are a known trend (Chiat & Roy, 2008; Yarian et al., 2021). However, this result presented an unexpected positive relationship, showing that children with a higher understanding of words yielded more errors in words and sentences when retelling the narrative. Although counterintuitive, this outcome is not the sole example of unexpected relations between vocabulary and expressive word knowledge and grammar. Illustratively, Le Normand et al., (2013) obtained a negative predictive relationship between the number of known lexical word types and MLU at ages 2 – 4 in a sample of 312 children. These less frequent findings show that the relationship between vocabulary and grammar is not yet entirely well understood (Brinchmann et al., 2018; Hoff et al., 2018). We speculate that our result could indicate that the more words children can understand the more linguistic pressure might exist to produce word forms that are not yet acquired well enough to be produced correctly. Consequently, this would yield word errors and non-sentences. One candidate of such words could be the verb forms, which are morphologically complex in Portuguese. Indeed, in our sample, around 53% of the errors occurred in the verbs' forms, where children used regularization of irregular verbs, singular suffixes instead of plural ones or/and wrong tenses.

The TENS was the unique metric of the narrative retelling that was correla-

ted with and predicted from item-based tests measures. This may have occurred for at least two reasons. One, because the TENS was the only narrative metric that did not depend directly on the narrative content and size, as the TNW, the MATTR and the MLU-w did. Also, the TENS combined sentences (non-sentences) production and words produced incorrectly (errors), being that these word errors did not alter the grammatical context, thus yielding a kind of lexical metric. So, the language units creating the TENS were the language units tapped by the two predictors, which confirms the essence of our second hypothesis.

Comparing the statistically significant relationships found in the TENS with those statistically non-significant relationships in the MLU-w, which also measures grammatical complexity, strengthens the idea that the production of errors in a narrative context is probably more discriminative than the MLU-w at the 4;11 to 6;0 age range. In fact, the MLU-w tends to not fully reflect the complexity of morphosyntactic skills when the overall level of language is already high (Frizelle et al., 2018), as was the case in the current study.

Our third hypothesis was validated. Only the item-based tests showed statistically significant increases in language levels from T1 to T2. While structured tests are designed to capture changes closely associated with age differences, the measures of grammatical complexity and lexical knowledge extracted from narratives tend to slow down after the ages of 5 – 6 years, requiring periods above one year to conceivably produce developmental significant differences (Frizelle et al., 2018; Rice et al., 2010). Accordingly, Santos et al. (2015) did not find statistically significant differences in the MLU-w between Group 2 (4;06-4;11 years) and Group 3 (5;00-5;05 years) in their Portuguese study, which led the authors to state that from the last year of kindergarten onwards, it is possible that children would be able to produce more complex sentences using fewer words. Thus, the MLU-w would no longer increase so steadily.

Following Voniati (2016), we would argue that the increase on narrative measures with age is likely to vary with experience and the language assessed. Thus, it is possible that in the Portuguese language, and for the age range of 5 to 6 years, item-based tests, compared to measures extracted from language samples, may be more sensitive to inter-individual variability. Nevertheless, although scarcely studied in Portuguese, it is important to highlight that our main objective was not to test the improvement of narrative skills across ages per se, but rather to understand if the item-based tests and the narrative task would or not have the same pattern of development across time, helping us in further understanding the relationships between these instruments.

As for the fourth hypothesis, it was confirmed that boys and girls generally presented the same levels of performance. This outcome is in line with reported outcomes of studies run with similar age groups regarding the MLU (e.g., Santos et al., 2015; Voniati, 2016), lexical diversity (e.g., Voniati et al., 2020), TNW (Hamdani et al., 2019), vocabulary (Lange et al., 2016), and item-based testing (e.g., Araújo et al., 2010). Recent evidence indicates that sex differences in langu-

age skills seem to be negligible and tend to disappear at the end of kindergarten (Etchell et al., 2018; Kidd & Donnelly, 2020).

Taken together, our results indicate that, when Portuguese kindergarten children have normal language development, each type of assessment tool – item-based tests and language samples from narrative retellings – would capture different kinds of language knowledge. The exception would be the narrative metric dealing with the production of errors which yielded statistically significant relations with, and was predicted by, item-based measures. This implies that at the 4;11 to 6;0 age range, a measure of the production of errors in narratives seem to be a good individual discriminative measure of language samples.

In conclusion, the results we observed in the current study are in alignment with the results obtained in most of the studies mentioned above, suggesting that item-based tests and language samples assess different aspects of language, particularly when the measures target different language units. Although there is divergent evidence regarding the relationship between the two types of assessment tools we examined, the results seem to be more coherent when the data is analyzed according to the processing unit – sentences or words. As the item-based tests used in this study involved massive amounts of word processing, the fact that we found a majority of statistically non-significant associations and predictions with measures extracted from the narratives appears to align with evidence indicating that the narrative skills and language skills measured by item-based tests constitute different language dimensions (Karlsen et al., 2021). While narrative retellings allow to access a child's core spontaneous language knowledge, providing detailed information about language difficulties, structured tests inform us about the expected skills profile for a given age group, in a specific dimension of language, that has a close dependence on learning opportunities from formal contexts such as school (Kopriva, 2000), which are not the same for everyone. The outcomes of the study also converge with the perspective that oral language presents a tendency of multidimensionality that is accentuated with the progression of the development, in which narrative skills and language skills measured by item-based tests appear to be separate constructs, particularly starting from the end of the kindergarten phase (Language and Reading Research Consortium [LARRC], 2015).

Consequently, the appropriate way to assess language skills and their development, particularly at the end of kindergarten, would be to use different types of instruments in a complementary way. A single approach in language assessment should not be the predominant system of practice, nor should we favor the idea that there is a language development representation that is the most appropriate one. Instead, the selection of the assessment process and tools should be guided by the objectives and the suitability for the target age.

In sum, besides showing that narrative samples and structured item-based tests tend to represent different language skills in kindergarten children, two relatively innovative trends emerged from the results: errors produced during the narrative assessments seem to be a good discriminative metric for the age range

studied, and testing language by units (words and sentences) might be an appropriate method to gather information about language development.

Limitations and Future Directions

One limitation of this study is its sample size of only 31 children, which could have impacted the results. A larger sample would have been beneficial for the statistical analyses that were carried out. Another aspect of this study that could be seen as a limitation is the small number of utterances and words per child. This probably stemmed from the narrative task, as the story was relatively small. The use of a longer story could elicit more words and utterances per child, which could have led to different results. Nevertheless, there is evidence that the number of utterances is not a decisive factor for the language samples' assessment quality (Miller et al., 2016; Pavelko & Owens, 2017) and the optimal length of the language samples might be dependent on the measure purpose (Guo & Eisenberg, 2015). Notwithstanding, it is important to state that our findings must be interpreted within the limits and conditions of the present study.

It would be interesting to replicate this study with younger Portuguese children, maintaining its longitudinal design over a longer period. Since language development occurs at a faster pace at earlier ages (Leitão, 2019), significant differences between moments of assessment and more significant associations between different instruments could arise. However, given the scarcity of literature regarding European Portuguese, this study is important in guiding the practice of Portuguese clinicians, considering that item-based tests are still widely used as a singular practice in the assessment of children's language skills.

Conflict of Interest Disclosure

The Authors declare no conflicts of interest.

Funding

The Authors declare no funding sources.

Research Ethics Statement

Procedures in this study adhere to ethical research policies and were approved by The University of Trás-os-Montes and Alto Douro (reference: DOC_FP. 19837-5035PA61968) and by the board of the children's schools. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. In addition, oral assent was obtained from each child at every testing session.

Data Availability Statement

Data can be obtained from the corresponding author upon reasonable request.

Authorship Details

Paula Alves: research concept and design, collection and/or assembly of data, data analysis and interpretation, writing the article, critical revision of the article, final approval of the article. Ana Paula Vale: research concept and design, collection and/or assembly of data, data analysis and interpretation, writing the article, critical revision of the article, final approval of the article.

References

Aguado, G. (1995). *El desarrollo del lenguaje de 0 a 3 años* [The development of language from 0 to 3 years]. CEPE.

Araújo, M. V. M., Marteleto, M. R. F., & Schoen-Ferreira, T. H. (2010). Avaliação do vocabulário receptivo de crianças pré-escolares [Assessment of kindergarteners' receptive vocabulary]. *Estudos de Psicologia (Campinas), 27*(2), 169–176. http://doi.org/10.1590/S0103-166X2010000200004

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T. H. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders, 43*(6), 498–510. https://doi.org/10.1016/j.jcomdis.2010.05.002

Beytollahi, S., Soleymani, Z., & Jalaie, S. (2020). The development of a new test for consecutive assessment of narrative skills in Iranian school-age children. *Iranian Journal of Medical Sciences, 45*(6), 425–433. https://doi.org/10.30476/ijms.2019.81984

Bishop, D. V. M. (2004). *Expression, reception and recall of narrative instrument*. Pearson.

Boerma, T., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2016). Narrative abilities of monolingual and bilingual children with and without language impairment: Implications for clinical practice. *International Journal of Language & Communication Disorders, 51*(6), 626–638. https://doi.org/10.1111/1460-6984.12234

Brinchmann, E. I., Braeken, J., & Lyster, S. A. H. (2018). Is there a direct relation between the development of vocabulary and grammar? *Developmental Science, 22*(1), e12709. https://doi.org/10.1111/desc.12709

Brown, R. (1973). A first language: The early stages. Harvard University Press.

Chiat, S., & Roy, P. (2008). Early phonological and sociocognitive skills as predictors of later language and social communication outcomes. *Journal of Child Psychology and Psychiatry, 49*(6), 635–645. https://doi.org/10.1111/j.

1469-7610.2008.01881.x

Cleave, P. L., Girolametto, L. E., Chen, X., & Johnson, C. J. (2010). Narrative abilities in monolingual and dual language learning children with specific language impairment. *Journal of Communication Disorders, 43*(6), 511–522. https://doi.org/10.1016/j.jcomdis.2010.05.005

Clegg, J., Hollis, C., Mawhood, L., & Rutter, M. (2005). Developmental language disorders –A follow-up in later adult life. Cognitive, language and psychosocial outcomes. *Journal of Child Psychology and Psychiatry, 46*(2), 128–149. https://doi.org/10.1111/j.1469-7610.2004.00342.x

Cronin, P., Reeve, R., McCabe, P., Viney, R., & Goodall, S. (2020). Academic achievement and productivity losses associated with speech, language and communication needs. *International Journal of Language & Communication Disorders, 55*(5), 734–750. https://doi.org/10.1111/1460-6984.12558

DeThorne, L. S., Johnson, B. W., & Loeb, J. W. (2005). A closer look at MLU: What does it really measure? *Clinical Linguistics & Phonetics, 19*(8), 635–648. https://doi.org/10.1080/02699200410001716165

Devescovi, A., & Caselli, M. C. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language and Communication Disorders, 42*(2), 187–208. https://doi.org/10.1080/13682820601030686

Diehm, E. A., Wood, C., Puhlman, J., & Callendar, M. (2020). Young children's narrative retell in response to static and animated stories. *International Journal of Language & Communication Disorders, 55*(3), 359–372. https://doi.org/10.1111/1460-6984.12523

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–4th edition*. Pearson Assessments.

Ebert, K. D., & Pham, G. (2017). Synthesizing information from language samples and standardized tests in school-age bilingual assessment. *Language, Speech, and Hearing Services in Schools, 48*(1), 42–55. https://doi.org/10.1044/2016_LSHSS-16-0007

Ebert, K. D., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools, 45*, 337–350. https://doi.org/10.1044/2014_LSHSS-14-0034

Etchell, A., Adhikari, A., Weinberg, L. S., Choo, A. L., Garnett, E. O., Chow, H. M., & Chang, S. (2018). A systematic literature review of sex differences in childhood language and brain development. *Neuropsychologia 114*, 19–31. https://doi.org/10.1016/j.neuropsychologia.2018.04.011

Ezeizabarrena, M. J., & Fernandez, I. (2018). Length of utterance, in morphemes or in words?: MLU3-w, a reliable measure of language development in early Basque. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.02265

Forrest, C. L., Gibson, J. L., Halligan, S. L., & St Clair, M. C. (2018). A lon-

gitudinal analysis of early language difficulty and peer problems on later emotional difficulties in adolescence: Evidence from the Millennium Cohort Study. *Autism & Developmental Language Impairments, 3*, 1–15. https://doi.org/10.1177/2396941518795392

Frizelle, P., Thompson, P. A., McDonald, D., & Bishop, D. V. M. (2018). Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language, 45*(5), 1–24. https://doi.org/10.1017/S0305000918000144

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U., & Walters, J. (2019). MAIN: Multilingual Assessment Instrument for Narratives – Revised. *ZAS Papers in Linguistics, 63*. https://doi.org/10.21248/zaspil.63.2019.516

Gazella, J., & Stockman, I. J. (2003). Children's story retelling under different modality and task conditions. *American Journal of Speech-Language Pathology, 12*(1), 61–72. https://doi.org/10.1044/1058-0360(2003/053)

George, D., & Mallery, P. (2016). *IBM SPSS statistics 23 step by step: A simple guide and reference*. Routledge.

Guiberson, M., & Ferris, K. P. (2019). Early language interventions for young dual language learners: A scoping review. *American Journal of Speech-Language Pathology, 28*(3), 945–963. https://doi.org/10.1044/2019_AJSLP-IDLL-18-0251

Guo, L. Y., & Eisenberg, S. (2015). Sample length affects the reliability of language sample measures in 3-year-olds: Evidence from parent-elicited conversational samples. *Language, Speech, and Hearing Services in Schools, 46*(2), 141–153. https://doi.org/10.1044/2015_LSHSS-14-0052

Hamdani, S. Z., Arshad, T., Tarar, S. A., & Kausar, R. (2019). Personal narrative skills of Urdu speaking preschoolers. *Narrative Inquiry, 29*(1), 50–81. https://doi.org/10.1075/ni.17063.ham

Hjetland, H., Lervåg, A., Lyster, S. H., Hagtvet, B. E., Hulme, C., & Melby-Lervåg, M. (2019). Pathways to reading comprehension: A longitudinal study from 4 to 9 years of age. *Journal of Educational Psychology, 111*(5), 751–763. http://dx.doi.org/10.1037/edu0000321

Hoff, E., Quinn, J. M., & Giguere, D. (2018). What explains the correlation between growth in vocabulary and grammar? New evidence from latent change score analyses of simultaneous bilingual development. *Developmental Science, 21*(2). https://doi.org/10.1111/desc.12536

Humphry, S., Heldsinger, S., & Dawkins, S. (2017). A two-stage assessment method for assessing oral language in early childhood. *Australian Council for Educational Research, 61*(2), 124–140. https://doi.org/10.1177/0004944117712777

Karlsen, J., Hjetland, H. N., Hagtvet, B. E., Braeken, J., & Melby-Lervåg, M. (2021). The concurrent and longitudinal relationship between narrative skills and other language skills in children. *First Language, 41*(5), 1–18. https://

doi.org/10.1177/0142723721995688

Kidd, E., & Donnelly, S. (2020). Individual differences in first language acquisition. *Annual Review of Linguistics, 6*(1), 319–340. https://doi.org/10.1146/annurev-linguistics-011619-030326

Klem, M., Melby-lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science, 18*(1), 146–154. https://doi.org/10.1111/desc.12202

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners.* The Council of Chief State School Officers.

Kuchirko, Y. (2019). On differences and deficits: A critique of the theoretical and methodological underpinnings of the word gap. *Journal of Early Childhood Literacy, 19*(4), 533–562. https://doi.org/10.1177/1468798417747029

Lange, B. P., Euler, H. A., & Zaretsky, E. (2016). Sex differences in language competence of 3- to 6-year-old children. *Applied Psycholinguistics, 37*(6), 1417–1438. https://doi.org/10.1017/S0142716415006244

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development, 86*(6), 1948–1965. https://doi.org/10.1111/cdev.12450

Law, J., Charlton, J., McKean, C., Beyer, F., Fernandez-Garcia, C., Mashayekhi, A., & Rush, R. (2018). *Parent-child reading to improve language development and school readiness: A systematic review and meta-analysis.* Newcastle University. https://eresearch.qmu.ac.uk/bitstream/handle/20.500.12289/9062/9062-Report.pdf?sequence=1&isAllowed=y

Le, H., Mensah, F., Eadie, P., McKean, C., Sciberras, E., Bavin, E. L., Reilly, S., & Gold L. (2020). Health-related quality of life of children with low language from early childhood to adolescence: Results from an Australian longitudinal population-based study. *Journal of Child Psychology and Psychiatry, 62*(3), 349–356. https://doi.org/10.1111/jcpp.13277

Le Normand, M. T., Moreno-Torres, I., Parisse, C., & Dellatolas, G. (2013). How do children acquire early grammar and build multiword utterances? A corpus study of French children aged 2 to 4. *Child Development, 84*(2), 647–661. https://doi.org/10.1111/j.1467-8624.2012.01873.x

Leitão, A. (2019). Desenvolvimento normal [Typical development]. In N. L. Antunes (Eds.), *Sentidos* [Senses] (pp. 33–70). Lua de Papel.

Lever, R., & Sénéchal, M. (2011). Discussing stories: On how a dialogic reading intervention improves kindergartners' oral narrative construction. *Journal of Experimental Child Psychology, 108*(1), 1–24. https://doi.org/10.1016/j.jecp.2010.07.002

Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research, 60*(8), 2185–2198. https://doi.org/10.1044/2017_JSLHR-L-15-0402

Luckman, C., Wagovich, S. A., Weber, C., Brown, B., Chang, S., Hall, N. E., & Ratner, N. B. (2020). Lexical diversity and lexical skills in children who stutter. *Journal of Fluency Disorders, 63,* 105747. https://doi.org/10.1016/j.jfludis.2020.105747

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed). Lawrence Erlbaum Associates.

MacWhinney, B., & Fromm, D. (2016). Child language data exchange system tools for clinical analysis. *Seminars in Speech and Language, 37*(2). http://dx.doi.org/10.1055/s-0036-1580743

Marôco, J., & Garcia-Marques, T. (2006). *Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas?* [What is Cronbach's alpha reliability? Old questions and modern solutions?] Laboratório de psicologia, 4(1), 65-90.

Melzi, G., & Schick, A. R. (2017). Language and literacy in the school years. In. J. B. Gleason, & N. B. Ratner (Eds.), *The development of language* (pp. 257–283). Pearson.

Mendes, A., Afonso, E., Lousada, M. & Andrade, F. (2014). *Teste de linguagem: Avaliação de linguagem pré-escolar* [Language test - Kindergarten language assessment]. Edubox SA.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2016). Using language sample analysis to assess spoken language production in adolescents. *Language, Speech, and Hearing Services in Schools, 47*(2), 99–112. https://doi.org/10.1044/2015_LSHSS-15-0051

Mimeau, C., Plourde, V., Ouellet, A., & Dionne, G. (2015). Comparison of measures of morphosyntactic complexity in French-speaking school-aged children. *First Language, 35*(2), 163–181. https://doi.org/10.1177/0142723715577320

Mouzaki, A., Spyropoulou, E., Ralli, A. M., Antoniou, F., Diamanti, V., & Papaioannou, S. (2020). The dimensionality of oral language ability: Evidence from young Greek children. *Journal of Speech, Language, and Hearing Research, 63*(8), 2732–2751. https://doi.org/10.1044/2020_JSLHR-19-00008

Nag, S., Snowling, M. J., & Mirkovic, J. (2018). The role of language production mechanisms in children's sentence repetition: Evidence from an inflectionally rich language. *Applied Psycholinguistics, 39*(2), 303–325. https://doi.org/10.1017/S0142716417000200

Norbury, C. F., & Bishop, D. V. M (2003). Narrative skills of children with communication impairments. *International Journal of Language and Communication Disorders, 38*(3), 287–313. https://doi.org/10.1080/136820310000108133

Owens, R. E. (2016). *Language development: An introduction.* Pearson Education Limited.

Owens, R. E., & Pavelko, S. L. (2017). Relationships among conversational language samples and norm-referenced test scores. *Clinical Archives of Communication Disorders, 2*(1), 43–50. https://doi.org/10.21849/cacd.2017.00052

Owens, R. E., & Pavelko, S. L. (2020). Sampling utterances and grammati-

cal analysis revised (SUGAR): Quantitative values for language sample analysis measures in 7-to 11-year-old children. *Language, Speech, and Hearing Services in Schools, 51*(3), 734–744. https://doi.org/10.1044/2020_LSHSS-19-00027

Paradis, J., Genesee, F., & Crago, M. (2021). *Dual language development and disorders: A handbook on bilingualism and second language learning* (3rd ed.). Paul Brookes Publishing

Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language, 25*(3), 365–376. https://doi.org/10.1177/0142723705059114

Paul, R., Norbury, C., & Gosse, C. (2018). *Language disorders from infancy through adolescence: Listening, speaking, reading, writing, and communicating.* Elsevier.

Pavelko, S. L., & Owens Jr, R. E. (2017). Sampling utterances and grammatical analysis revised (SUGAR): New normative values for language sample analysis measures. *Language, Speech, and Hearing Services in Schools, 48*(3), 197–215. https://doi.org/10.1044/2017_LSHSS-17-0022

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research, 53*(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Rinaldi, P., Pasqualetti, P., Volterra, V., & Caselli, M. C. (2021). Gender differences in early stages of language development. Some evidence and possible explanations. *Journal of Neuroscience Research, 101*(5), 643–653. https://doi.org/10.1002/jnr.24914

Roy, P., & Chiat, S. (2013). Teasing apart disadvantage from disorder: The case of poor language. In C. R. Marshall (Ed.), *Current issues in developmental disorders* (pp. 125–150). Psychology Press.

Roy, P., Chiat, S., & Dodd, B. (2014). *Language and socioeconomic disadvantage: From research to practice*. City University London.

Santos, M. E., Lynce, S., Carvalho, S., Cacela, M., & Mineiro, A. (2015). Extensão média do enunciado-palavras em crianças de 4 e 5 anos com desenvolvimento típico da linguagem [Mean length of utterance-words in 4 and 5-year-old children with typical language development]. *CEFAC, 17*(4), 1143–1151. https://doi.org/10.1590/1982-021620151741315

Sapage, S., Cruz-Santos, A., & Abreu, P. E. (2019). O programa CLAN: Transcrição e análise de narrativas orais de crianças [The CLAN program: Transcription and analysis of children's oral narratives]. In A. J. Osório, M. J. Gomes & A. L. Valente (Eds.), *Challenges 2019: Desafios da inteligência artificial* [Artificial intelligence challenges] (pp. 455–466). Competence centre of Communication and Information Technologies in Education, Min-

ho University. https://www.nonio.uminho.pt/wp-content/uploads/2020/09/atas_ch2019_full.pdf

Schoon, I., Parsons, S., Rush, R., & Law, J. (2010). Children's language ability and psychosocial development: A 29-year follow-up study. *Pediatrics, 126*(1), e73–e80. https://doi.org/10.1542/peds.2009-3282

Silva, I. L., Marques, L., Mata, L., & Rosa, M. (2016). *Orientações curriculares para a educação pré-escolar* [Curricular guidelines for pre-school education]. Ministério da Educação/Direção-Geral da Educação. https://www.dge.mec.pt/sites/default/files/Noticias_Imagens/ocepe_abril2016.pdf

Southwood, F., & Russell, A. F. (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research, 47*(2), 366–376. https://doi.org/10.1044/1092-4388(2004/030)

Stipdonk, L. W., Dudink, J., Reiss, I. K., & Franken, M. C. J. (2020). Does a narrative retelling task improve the assessment of language proficiency in school-aged children born very preterm? *Clinical Linguistics & Phonetics, 34*(12), 1112–1129. https://doi.org/10.1080/02699206.2020.1720824

Stockman, I. J. (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools, 27*(4), 355–366. https://doi.org/10.1044/0161-1461.2704.355

Tomas, E., & Dorofeeva, S. (2019). Mean length of utterance and other quantitative measures of spontaneous speech in Russian-speaking children. *Journal of Speech, Language, and Hearing Research, 62*, 4483–4496. https://doi.org/10.1044/2019_JSLHR-L-18-0339

UK Department for Education (2020). *Nuffield early language intervention programme (NELI)*. https://www.gov.uk/government/publications/catch-up-premium-coronavirus-covid-19

Ukrainetz, T. A., & Blomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy, 18*(1), 59–78. https://doi.org/10.1191/0265659002ct227oa

Vicente, S. G., Sousa, A. S., & Silva, M. A. (2011a). *Adapted version of Peabody Picture Vocabulary Test-4 from Dunn & Dunn* (2007) [Master thesis, University of Porto].

Vicente, S. G., Sousa, A. S., & Silva, M. A. (2011b). *European Portuguese material for receptive vocabulary assessment [working research version constructed based on the PPVT-4 (Form B) of Dunn, & Dunn, 2007]*. Unpublished Material, Centre of Psychology, Faculty of Psychology and Educational Sciences, University of Porto, Porto, Portugal.

Voniati, L. (2016). Mean length of utterance in Cypriot Greek-speaking children. *Journal of Greek Linguistics, 16*(1), 117–140. https://doi.org/10.1163/15699846-01601002

Voniati, L., Tafiadis, D., Armostis, S., Kosma, E. I., & Chronopoulos, S. K.

(2020). Lexical diversity in Cypriot-Greek-speaking toddlers: A preliminary longitudinal study. *Folia Phoniatrica et Logopaedica, 73*(4), 1–12. https://doi.org/10.1159/000507621

West, G., Snowling, M. J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H., & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: Evidence from a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry, 62*(12), 1425–1434. https://doi.org/10.1111/jcpp.13415

Westerveld, M. F., & Gillon, G. T. (2010). Profiling oral narrative ability in young school-aged children. *International Journal of Speech-Language Pathology, 12*(3), 178–189. https://doi.org/ 10.3109/17549500903194125

Westerveld, M. F., Gillon, G. T., & Boyd, L. (2012). Evaluating the clinical utility of the profile of oral narrative ability for 4-year-old children. International *Journal of Speech-Language Pathology, 14*(2), 130–140. https://doi.org/10.3109/17549507.2011.632025

Westerveld, M. F., & Vidler, K. (2016). Spoken language samples of Australian children in conversation, narration and exposition. *International Journal of Speech-Language Pathology, 18*(3), 288–298. https://doi.org/10.3109/17549507.2016.1159332

Wu, S. Y. (2020). Mean length of utterance among Mandarin-speaking children with and without DLD. *Child Language Teaching and Therapy, 36*(3), 165–179. https://doi.org/10.1177/0265659020945366

Yarian, M., Washington, K. N., Spencer, C. E., Vannest, J., & Crowe, K. (2021). Exploring predictors of expressive grammar across different assessment tasks in preschoolers with or without DLD. *Communication Disorders Quarterly, 42*(2), 111–121. https://doi.org/10.1177/1525740119868238

Zhang, X., Qin, F., Chen, Z., Gao, L., Qiu, G., & Lu, S. (2020). Fast screening for children's developmental language disorders via comprehensive speech ability evaluation – Using a novel deep learning framework. *Annals of Translational Medicine, 8*(11). http://dx.doi.org/10.21037/atm-19-3097

Appendix

Image 1 of the Narrative Retelling in Portuguese Task.