CITY UNIVERSITY LONDON

EST 1894

City Research Online

Permanent City Research Online URL: http://openaccess.city.ac.uk/4805/

Sentence repetition: What does the task measure?

Kamila Polišenská†‡, Shula Chiat† and Penny Roy†

†Language and Communication Science, City University London, London, UK

Research Report

Running head: Sentence repetition: What does the task measure?

Abstract

Background: Sentence repetition is gaining increasing attention as a source of information about children's sentence-level abilities in clinical assessment, and as a clinical marker of Specific Language Impairment. However, it is widely debated what the task is testing and therefore how informative it is.

Aims: The study aims to: i) evaluate the effects of different types of long-term linguistic knowledge on immediate recall, ii) assess age sensitivity of repetition tasks designed to evaluate these effects and iii) establish if the effects are similar across typologically different languages. The study also considers the implications of the findings for the use of sentence repetition as a research and clinical assessment tool.

Methods & Procedures: Participants were 50 English-speaking and 50 Czech-speaking typically developing 4-5-year-olds. Children's ability to recall sequences of items was compared in seven linguistic conditions, ranging from fully well-formed sentences to sequences of nonwords. In each condition, children repeated blocks of successively longer stimuli to establish their span.

Outcomes & Results: Results showed significant but differential effects of all linguistic factors in both languages. While syntactic violations and presence of nonwords dramatically reduced children's span, semantic implausibility and the removal of sentence prosody played a significant but much smaller role. Familiarity of function words was more important than

familiarity of content words. The effects of different linguistic factors on spans were the same for both languages and did not change between 4 and 5 years, although average spans increased over this age range.

Conclusions & Implications: Children's ability to repeat sentences is more dependent on their familiarity with morphosyntax and lexical phonology than semantics or prosody, with function words of particular importance. Findings have implications for the use of recall in clinical assessment and as a research tool.

What is already known on this subject

Repetition tasks often feature in standardised language assessments and have emerged as reliable clinical markers of language impairment in children across different populations and languages. While it is clear that sentence repetition draws on children's linguistic knowledge, it is not clear how different types of linguistic knowledge contribute to children's performance, and hence what sentence repetition tasks tell us about language abilities and language impairment in children.

What this study adds

The study provides further evidence that immediate verbal repetition is highly sensitive to linguistic structures present in the stimuli. It adds to existing evidence by demonstrating that lexical phonology and morphosyntax play a key role in typically developing children's immediate recall capacity. Findings clarify what sentence repetition tasks test and the particular contribution they can make to language assessment of children.

1. Introduction

Repetition tasks are increasingly used in language assessment and in research, but there is little agreement on the tasks' underlying mechanisms and hence on their clinical informativeness. For a long time, research has focused on nonword repetition (for a review see Coady & Evans, 2008), but more recently sentence repetition has been highlighted as a potential clinical marker of Specific Language Impairment (SLI) in English and other languages (Conti-Ramsden, Botting, & Faragher, 2001; Everitt, Hannaford, & Conti-Ramsden, 2013; Riches, 2012; Seeff-Gabriel, Chiat, & Dodd, 2010; Stokes, Wong, Fletcher, & Leonard, 2006). Sentence recall (sentence repetition/elicited imitation) has also been used as a method for exploring different aspects of syntactic development (e.g. Kidd, Brandt, Lieven, & Tomasello, 2007). Yet the nature of the mechanisms underlying sentence repetition and the contribution of linguistic knowledge remain unclear. The aim of this study is to provide better understanding of the processes underlying immediate repetition of sentences in order to establish what poor performance on these tasks actually reveals about children's language. This in turn has implications for models of verbal short-term memory (STM), and for 'limited capacity' and 'linguistic theories' of SLI (Leonard, Ellis, Miller, Francis, & Tomblin, 2007; Mainela-Arnold, Evans, & Coady, 2010; Mainela-Arnold; Misra, Miller, Poll, & Park, 2012).

1.1.   The impact of long-term linguistic knowledge on immediate repetition

Linguistic knowledge has previously been shown to influence immediate verbal repetition. For example, capacity is greater for lists containing words of high rather than low frequency (Mainela-Arnold & Evans, 2005), and imageable/concrete words are recalled better than abstract words (Walker & Hulme, 1999). Even recall of nonwords, which by definition are not stored in long-term memory (LTM), shows effects of linguistic factors. Children's

nonword repetition performance is influenced by phonological properties including prosody (Roy & Chiat, 2004), and syllable complexity (Marshall & van der Lely, 2009), and by lexical phonological properties (e.g. Gathercole, Frankish, Pickering, & Peaker, 1999; Metsala & Chisholm, 2010). These robust findings demonstrate that verbatim recall draws on and benefits from participants' long-term linguistic knowledge, even when stimuli (nonwords) are specifically designed to minimise potential support from long-term knowledge.

In response to evidence of LTM influence on immediate recall, Baddeley (2000) added the 'episodic buffer' to his working memory model. This additional component accommodates the contribution of LTM from word up to sentence level. However, as Allen and Baddeley (2009) and Baddeley, Hitch and Allen (2009) acknowledge, the model still requires specification regarding the precise nature of the contribution of linguistic domains. Other theories of recall argue that linguistic knowledge and memory are intertwined in immediate recall tasks (Acheson & MacDonald, 2009), and that STM cannot be separated from language comprehension and production processes. This raises questions about which components of comprehension and production are activated in sentence recall. Our study set out to address the acknowledged gap in the Baddeley model, and in so doing, to throw more light on the aspects of comprehension and production that are involved in immediate recall according to Acheson and MacDonald.

Our approach builds on a landmark study by Miller and Isard (1963) which systematically manipulated the semantic and syntactic well-formedness of sentences in order to investigate the effects of linguistic domains on adults' ability to recall sentences. This study presented triplet sentences such as: i) A witness signed the official legal document (Grammatical Sentence), ii) A witness appraised the shocking company dragon (Anomalous Sentence), iii) A legal glittering the exposed picnic knight (Ungrammatical Sentence). Miller and Isard found that the absence of syntactic structure hindered recall more than semantic

implausibility. This throws some light on the contribution of language domains to sentence-level processing in a mature language system, indicating the extent to which adults draw on syntactic knowledge to support recall. However, the outcome measure in this study was the number of trials that adults needed for perfect recall, making this a test of learning rather than immediate recall. It therefore has little bearing on the immediate sentence recall tasks used in clinical assessment of children.

Few studies have examined the contribution of linguistic domains to recall in children. Bonvillian, Raeburn and Horan (1979) investigated the effect of intonation on imitation of short and long sentences in children, finding that imitation of sentences with normal intonation was superior to imitation of sentences with no intonation, although this effect was only found in longer sentences. Other studies have focused on the distinction between content words (CWs) and function words (FWs). Scholes (1970) compared children's recall of well-formed sentences and ungrammatical sentences, scoring the number of correctly repeated CWs and FWs. FWs were more prone to omission than CWs, and younger children omitted FWs regardless of the grammaticality of the string in which they occurred. As age increased, FW omission decreased in the grammatical sentences, disappearing entirely in adult performance. Gerken, Landau and Remez (1990) investigated the effects of replacing inflections and FWs with nonwords, creating strings with English or nonsense function morphemes and English or nonsense content words. Children tended to retain strongly stressed elements and omit weakly stressed elements, particularly when the elements carried morphosyntactic information.

These recall studies with children have largely focused on a single linguistic factor (e.g. lexical status of CWs/FWs, prosody, syntax) and used the number of words correctly recalled in the sentences as the outcome measure. The current study draws together linguistic factors that have been found separately to influence recall in either adults or children in order

to evaluate their relative importance, and uses span as a measure of recall beyond the single word level. No previous study has, to our knowledge, manipulated the full range of linguistic factors we have examined to provide a more complete picture of how specific language domains contribute to children's immediate recall of sentences.

1.2.   The present study

While there has been a great deal of research on repetition of single unrelated items (words and nonwords), the repetition of words linked at a sentence level has received less attention. Repetition at sentence level provides a better platform for exploring the influence on immediate recall of long-term linguistic knowledge beyond the lexical and sublexical properties of single items. The present study exploited this potential to address the following questions:

i) What are the relative effects of different types of long-term linguistic knowledge on span?

ii) Is repetition performance for different types of long-term linguistic knowledge age sensitive?

iii) Are effects of long-term linguistic knowledge on repetition performance similar across typologically different languages?

Our study evaluated the role of linguistic knowledge in sentence recall by systematically deconstructing the linguistic representations present in sentences and assessing the extent to which each type of representation affects repetition capacity. Two levels of sentence representation are relatively independent of all others and can therefore be manipulated independently of others: semantic plausibility of sentences can be violated without affecting words, grammar, or prosody (as in the anomalous sentence 'She sang us a

kettle'; likewise, sentence prosody can be eliminated without affecting words, grammar, or semantics (as in 'she, sent, us, a, letter' presented with the prosody of a list). It was therefore possible to isolate the effects of semantic plausibility and sentence prosody on recall capacity. In contrast, lexicality, grammaticality and semantics are necessarily intertwined, making it impossible to completely separate these out. Syntax conveys the grammatical roles of words/phrases, so it is not possible to disrupt syntax without disrupting syntactic and hence semantic relations; nor is it possible to substitute nonwords for real words without disrupting both syntax and semantics. However, languages deploy a limited range of formal devices to convey meaning and meaning relations, and it is possible to isolate each of these and thereby investigate their relative effect on recall capacity. The formal devices to which we refer are:

- Content words, an open class of phonological forms that convey the key conceptual information in a sentence (events and states; people and things that participate in these; timing of events; and manner of events).

- Word order, which conveys meaning relations between content words.

- Function words, a closed class of phonological forms that typically serve to modify content words and/or indicate relations between these, and often have distinct phonological characteristics (e.g. in English, most function words are monosyllabic and have reduced forms).

- Inflections, also a closed class of phonological forms that serve to modify content words and/or indicate relations between these, but unlike function words, are phonologically incorporated into the phonology of the content word they modify.

    In this study, we manipulated semantic plausibility, prosody, word order, content words and function words to determine which aspects of sentence input are more or less crucial for immediate recall. These manipulations resulted in a spectrum of linguistic

conditions from entirely well-formed sentences at one extreme, to strings of nonwords at the other. Note that we did not manipulate inflections, which are limited and play a marginal role in marking syntactic and semantic relations in English (see below).

We conducted our study with typically developing children aged 4-5 years in two languages, Czech and English, taking advantage of the fact that the researcher is a native speaker of Czech. This afforded the opportunity to compare effects of our target factors in two languages differing typologically in the devices that mark grammatical roles and are crucial for sentence processing. In contrast to English, which has a sparse morphology and is heavily reliant on word order and function words, Czech has a rich morphology and is heavily reliant on inflections (which were not manipulated in this study), with relatively free word order (which was manipulated). Czech word forms typically consist of a stem plus an ending marking multiple grammatical categories and relations. All nouns carry grammatical gender (masculine, feminine, or neuter), and are declined for both number (singular, plural) and case (nominative, genitive, dative, accusative, vocative, locative, and instrumental). In addition, masculine paradigms are marked for animacy. Within a noun phrase, a noun or pronoun and all modifiers show agreement for case, gender and number and the presence of case markers on nouns, pronouns, and adjectives combines with a freer word order than English. Czech is a pro-drop language and pronominal subjects are usually omitted unless special emphasis is required. By manipulating the same set of variables in the English and Czech stimuli, we were able to investigate if, and how far, memory span is influenced by each language's specific morphosyntactic devices.

In order to investigate the effects of our target variables, our study used the methodology of 'span' tasks which establish the maximum number of words that participants are able to repeat correctly (their 'span'). Span is typically used as a measure of recall for strings of items (digits or words). Our novel use of a span task with sentences instead of

single items enabled us i) to quantify the impact of each type of linguistic information on immediate recall capacity in children, and ii) to determine whether the impact of each type of linguistic information is affected by differences between languages. Using span as the outcome measure had the benefit of targeting children's threshold for recall and avoided likely floor effects (for ungrammatical sentences) and ceiling effects (for well-formed sentences) if we had sought to match item length across conditions. At a theoretical level, using a task that is widely treated as a measure of memory capacity enables us to determine the extent to which memory capacity depends on linguistic knowledge, with implications for the interpretation of children's difficulties with repetition tasks, and relations between verbal short-term memory and language deficits in children.

2. Method

2.1. Participants

One hundred typically developing children participated in this study: 50 English-speaking (age range 4;0-5;11, mean age 4;11, SD = 6 months) and 50 Czech-speaking children (age range 4;1-5;11, mean age 4;11, SD = 6 months). The participants were recruited via nurseries and primary schools in London and the Kroměříž region of the Czech Republic. Formal ethical approval for the study was granted by City University Research Ethics Committee. Children were only included if they had never been referred for speech and language therapy; had no known hearing loss or neurological problems; spoke English or Czech as a first language; and if their parents gave signed consent and children provided their assent. In total, 53 English-speaking and 53 Czech-speaking children were recruited, but six participants had to be excluded: three children did not finish the task and three children had unintelligible speech. The non-compliance rate was very low (3%); from 103 children who attempted the task, 100 children completed it.

Children's general language ability was assessed by a receptive vocabulary test. In the English sample, the British Picture Vocabulary Scale (BPVS-II, Dunn, Dunn, Whetton, & Burley, 1997) was used. Due to the lack of a standardised vocabulary test in Czech, the BPVS-II was used and adapted when necessary. Table 1 shows that the samples were well matched for age, so raw scores were used. No significant differences were found between Czech- and English-speaking children in their BPVS raw scores, suggesting that the samples were broadly comparable in their receptive vocabulary knowledge. The mean scores also show that vocabulary knowledge increased with age in both language groups, confirming the age-sensitivity of the task.

Table 1. Overview of participants' age and receptive vocabulary scores according to language and age group.

	4-year-olds		5-year-olds	
	English	Czech	English	Czech
Mean Age (SD)	54.04 (2.84)	54.69 (2.80)	66.22 (3.91)	65.33 (3.92)
Age differences	t(51) = -.85, p = .40		t(45) = .77, p = .44	
BPVS (SD)	55.11 (10.69)	56.62 (20.74)	63.74 (13.94)	73.50 (17.77)
BPVS differences	t(37.1*) = -.33, p = .74		t(45) = -1.87, p = .07	

Note: *Homogeneity of variance not assumed.

2.2. Experimental task

The study comprised seven conditions, with the fully well-formed sentences in the first condition providing the foundation for systematic manipulation of linguistic factors in the other six. For each condition, eight blocks of four sentences were constructed, starting with two-word stimuli in the first block (e.g. 'She cooks' in English and the Czech equivalent 'Ona

vaří'), and successively increasing by one item up to the final block of nine-word stimuli (e.g., 'The young dancer was looking at the shiny mirror' and the Czech example 'Ten mladý herec se díval do toho velkého zrcadla' where 'shiny' was replaced with 'big'). Table 2 provides an overview of the seven conditions with examples of targets at length five words.

Table 2. Experimental conditions with examples

Condition	Example
A: Grammatical sentence	He sent us a letter
B: Sentence with list prosody	he, sent, us, a, letter
C: Semantically implausible sentence	He sang us a kettle
D: Ungrammatical sentence (random word order)	A sent he letter us
E: Pseudosentence with content words replaced by nonwords	*He /fɪnt/ us a /lɔpə/*
F: Pseudosentence with function words replaced by nonwords	/vi/ sent /əʃ ʊ/ letter
G: Pseudosentence with all words replaced by nonwords	*/vi fɪnt əʃ ʊ lɔpə/*

2.2.1. Experimental conditions

Since the main aim of the study was to determine the contribution of different linguistic domains rather than different syntactic structures, the well-formed foundation sentences in condition A were simple, main clauses. These were produced with neutral prosody. The stimuli in English and Czech were matched in number of words, balance of CWs and FWs and semantics of words and sentences, but diverged in aspects of lexical phonology and morphosyntax reflecting characteristics of each language.

Lexical items were selected to be appropriate for the target age group (4-5-year-olds), and were of high familiarity and imageability. For English, information about familiarity, imageability and age of acquisition of lexical items was obtained from the MRC

Psycholinguistic Database (Coltheart, 1981). In the absence of a database providing this information for Czech, the lexical items were given to Czech native speakers who were linguists and nursery teachers to judge their age appropriateness. To avoid potential difficulties in speech output, the words/nonwords in English were no more than disyllabic, carried trochaic stress, and contained no more than two adjacent consonants within a single syllable. Czech words/nonwords contained no more than three adjacent consonants within a single syllable, were trisyllabic or less and carried trochaic or dactylic stress, reflecting the characteristic structure of Czech words.

Stimuli in the other linguistic conditions were created from the grammatical sentences in condition A as described below:

B: Grammatical sentences with list prosody were a counterpart to the grammatical sentences in condition A, with items presented using list rather than sentence prosody.

C: Semantically implausible sentences replaced CWs in condition A with lexical items that were phonologically matched but resulted in semantic implausibility. Implausibility arose from violations of animacy and concreteness restrictions on verb arguments, e.g. sock brushed its eye, flower was sick, she snows, wash the voice, and incompatibility of adjective-noun combinations, e.g. sweet salt, sunny hammer. One can presume that the children had never encountered these particular word combinations before and that frequency of co-occurrence of constituent items was low relative to the plausible sentences. The prosodic structure of the original semantically plausible sentences was preserved.

D: Ungrammatical sentences disrupted syntax by altering the order of words in condition A. Although Czech has relatively free phrase order, word order is fixed within phrases and therefore can be violated. As well as allowing direct comparison between Czech and English, the advantage of investigating violation of word order (as opposed to violating

grammatical morphology in Czech) was that no phonological material was lost or altered, and that the grammatical and ungrammatical sentences contained exactly the same lexical items and affixes. The ungrammatical sentences were produced with stress appropriate for content and function words within a sentence.

E: Pseudosentences with pseudocontent words were created by replacing items in the content word slots with phonologically matched nonwords, preserving the morphosyntactic frame and prosodic properties of the original sentence, as in He /fɪnt/ us a /lɔpə/. Nonwords and real words were matched for number of syllables and syllabic structure, e.g. 'sent' and /fɪnt/ both with a CVCC structure. This resulted in many cases in rhyming, but there were also examples that words/nonwords did not rhyme, e.g. /sɔɪ/ for 'buy'.

F: Pseudosentences with pseudofunction words were created by replacing FWs with phonologically matched nonwords. The sequence of content words and prosody of the original sentences were retained, with CWs stressed and pseudofunction words unstressed, as in /vi/ sent /əʃʊ/ letter.

G: Pseudosentences made up of nonwords were created by replacing all words from condition A with phonologically matched nonwords. Prosody of items matched their prosody in the original sentences, as in /vi fɪnt əʃ ʊ lɔpə /.

     The stimuli were recorded in a quiet room by a native speaker of either British English or Czech. The order of conditions was counterbalanced.

2.2.2. Span task

     Span tasks are well established as a method for measuring STM capacity, most typically using digits and sometimes words/nonwords. The present experiment followed the procedure used in the Working Memory Test Battery for Children (WMTB-C; Pickering &

Gathercole, 2001), but with certain adjustments due to the number of conditions in the current experiment. Four (rather than six) trials were presented at each length in each condition. To pass at a particular length, children had to repeat three out of four trials correctly. If the participant failed to repeat three out of four stimuli at a particular length, testing moved backwards; if the previous block had already been passed, testing stopped. If the participant succeeded at a particular length, testing progressed through successive lengths until they were unable to repeat three out of four of the stimuli in a block.

The span credited to the child in each condition was the highest target length at which the participant could successfully repeat at least three of the four targets. To allow for more fine-grained scoring, a half point was awarded when the participant correctly produced two stimuli at the next highest/first target length.

The span task was embedded in a game using Microsoft PowerPoint, 2003. Each condition was represented by one 'playing field' showing a 'snake' of 32 coloured ovals containing sound files for the 32 stimuli in that condition (4 at each of 8 lengths). After each playing field was completed (i.e. when span for that condition was established), a short animation was presented on a new slide and children could choose a sticker. The animation was intended to reward the participants for completing each playing field and provide motivation to attempt the next playing field (i.e. next experimental condition). It also served as a filler, giving participants the opportunity to discuss events on the screen and break the monotony of repetition.

2.3. Procedure

Children were tested in one session which lasted on average 25 minutes. The receptive vocabulary test was administered first. Before the stimuli were presented, participants were told that the game consisted of listening to words and copying what they heard. Following

four practice trials, the seven playing fields/conditions were presented. Each condition started

with a sentence/string length expected to be well within the participant's capacity based on

age and findings of pilot studies carried out in both languages (details available from the first

author), and moved forward, backward or stopped depending on the child's performance as

described above. In the nature of a span task, the number of stimuli each child received

varied. Administration was manually controlled by the researcher and each stimulus was only

played once. The participant's response to each stimulus was recorded on a score sheet and

on a Marantz Professional PMD620 digital recorder.

Throughout the task, participants were verbally praised regardless of their

performance and encouraged to continue the task until stimuli from all of the experimental

conditions had been administered. At the end of the session, participants were thanked for

taking part.

2.4. Scoring criteria

Repetition of each target was scored as correct/incorrect as a whole. In lexical

conditions, repetition was correct if i) all words were present in the correct order, ii) all

syllables of the words were present in the correct order and iii) all inflections were present

(e.g. 3.sg. present tense -s, plural -s, past tense -ed). As phrasal order is flexible in Czech,

word order changes were allowed in the Czech conditions provided grammatical and meaning

relations of the target were preserved. In non-lexical conditions, responses were correct if i)

all items were repeated in the correct order and ii) all syllables were repeated in the correct

order. Since precise repetition of nonwords was not a key issue, some allowances were made

for the syllables within nonwords. These were considered to be correct if i) at least one

consonant from a syllable onset was present, e.g. target /brul/ – responses such as [rul] or

[bul] would count as correct ii) the nucleus vowel was identical or differed in no more than

one feature (front vs. back, high vs. low, tense vs. lax), e.g. target /lɪm/ - [ləm] accepted as correct.

3. Results

3.1. Reliability

Following Hallgren (2012), inter-rater reliability was evaluated using the intra-class correlation coefficient (ICC). Cicchetti (1994) described values between .60 and .74 as good, and values between .75 and 1.0 as excellent. A random ten percent of the samples (the results of 5 English-speaking children and 5 Czech-speaking children) were selected for independent raters to score. Six conditions in English showed excellent agreement (ICC between .79 and 1), and one condition showed good agreement (ICC = .72). In Czech, six conditions showed excellent agreement (ICC between .98 and .83) and one condition showed good agreement (ICC=.73). Overall these results showed the scoring system was reliable and fit for purpose.

3.2. Effects of linguistic conditions

The key objectives of this study were (i) to quantify the impact that each type of linguistic information had on immediate recall in children, and (ii) to determine if the repetition task was age sensitive, (iii) to determine whether the impact of linguistic information was affected by morphosyntactic differences between languages. Table 3 shows the mean spans and standard deviations for each of the seven experimental conditions by language and age group. As can be seen in Figure 1, although the mean maximum spans for the Czech sample were marginally lower than the English sample across all conditions, the pattern of results was almost identical for the two language groups. The linguistically well-formed items yielded the highest span in both English and Czech, while the least linguistically well-formed items yielded the lowest span.

Table 3. Mean spans and standard deviations across experimental conditions by language and age (English in white, Czech in grey).

| | English | | | | | | Czech | | | | | |
| | 4-yr-olds | | 5-yr-olds | | Whole sample | | 4-yr-olds | | 5-yr-olds | | Whole sample | |
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	7.61	1.33	8.48	1.14	8.01	1.31	7.44	1.03	7.73	1.14	7.58	1.08
B	7.09	1.39	7.98	1.19	7.50	1.37	6.90	1.30	7.10	1.25	7.00	1.27
C	6.69	1.46	7.48	1.34	7.05	1.45	6.56	1.03	6.94	1.24	6.74	1.14
D	4.13	1.10	4.61	0.81	4.35	1.00	3.77	0.64	4.04	0.71	3.90	0.68
E	4.02	1.03	4.74	0.95	4.35	1.05	3.04	0.77	3.88	0.95	3.44	0.95
F	2.85	0.76	3.20	0.96	3.01	0.87	2.17	0.79	2.79	1.14	2.47	1.01
G	2.63	0.73	3.09	0.81	2.84	0.79	2.35	0.75	2.75	0.69	2.54	0.74

Figure 1. Mean span for experimental conditions, ranging from grammatical sentences to nonlexical strings in English- and Czech-speaking children. Error bars are standard deviations.
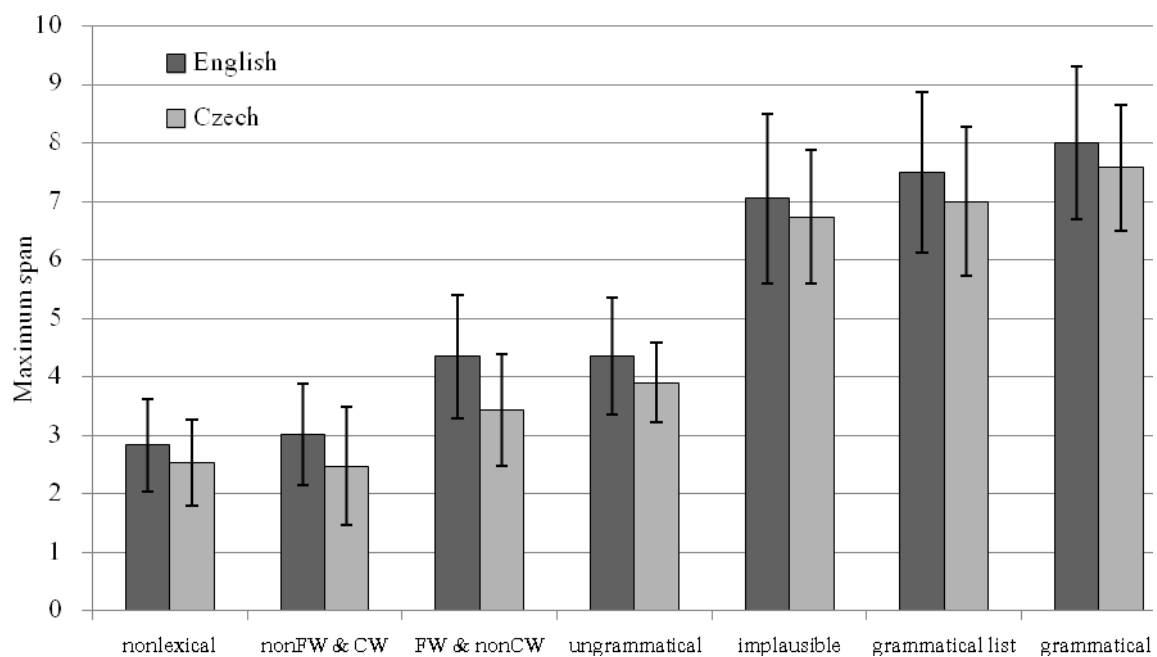
Due to the typological differences between the languages, separate mixed ANOVAs were run for English and Czech. The first set of 2 x 7 ANOVAs investigated the effects Age group and Condition, separately for English and Czech. In English, the ANOVA revealed a highly significant main effect of Age ($F(1, 48) = 9.30$, $p = .004$, $\eta^2 = .16$), Condition ($F(4.42, 212.23) = 311.70$, $p < .001$, $\eta^2 = .87$), with no interaction between Age*Condition ($F(4.42, 212.23) = .79$, $p = .546$, $\eta^2 = .02$). The pattern of results was the same for Czech: a highly significant main effect of Age ($F(1, 48) = 4.57$, $p = .038$, $\eta^2 = .09$) and Condition ($F(4.35, 217.3) = 452.02$, $p < .001$, $\eta^2 = .90$), with no interaction between Age*Condition ($F(4.35, 217.3) = 1.16$, $p = .330$, $\eta^2 = .02$).

Not all the post hoc comparisons between conditions were of theoretical interest; the aim was rather to establish if there was an effect of grammaticality, lexicality, plausibility and prosody. This was explored through selected comparisons. The effect of semantic plausibility was assessed by comparing the span for semantically plausible and semantically implausible sentences (A vs. C); grammaticality was assessed by comparing the span for grammatical and ungrammatical sentences (A vs. D); lexicality was assessed by comparing the span for lists of words in random order and lists of nonwords (D vs. G); and prosody was assessed through a comparison of grammatical sentences with and without prosody (A vs. B). The effect of lexical status of function words and content words was assessed by comparing hybrid strings with real FWs and non-CWs vs. non-FWs and real CW (E vs. F).

Table 4 shows the mean differences in span and post-hoc analyses (Bonferroni) for the target comparisons together with effect sizes (Cohen's d) of the differences. Effect sizes were interpreted in line with Cohen's (1988) guidelines as d of .2 is 'small', .5 'medium' and .8 is 'large' effect. These revealed that each language domain had a significant effect on immediate recall. The largest difference was found for grammaticality, with children attaining

a mean span over 3.5 words longer for grammatical than ungrammatical word order. The advantage for other factors was notably smaller, from around 1.5 for lexicality, and just under 1 word for plausibility, to about 0.5 for prosody. This stepwise pattern was found for both English and Czech. Post-hoc analyses also revealed that the span for pseudosentences with real FWs but nonwords replacing CWs (FW + non-CW) was significantly longer than the span for pseudosentences with real CWs but nonwords replacing FWs (non-FW + CW). This effect was observed in both languages (with an advantage of over one word for English, and about one word for Czech), even though the Czech stimuli provided more inflections and therefore more grammatical information in the condition with real FWs+ nonlexical CWs than the English stimuli.

Table 4. Mean differences in span and results of post hoc comparisons (English/Czech)

Domains	Mean difference	SE	p-value	Cohen's d
	EN/CZ	EN/CZ	EN/CZ	EN/CZ
Prosody	0.51/0.58	.15/.14	.038/ .004	0.38/0.49
Plausibility	0.96/0.84	.17/.12	<.001	0.70/0.76
Lexical status of CWs vs. FWs	1.36/0.97	.15/.14	<.001	1.39/0.99
Lexicality	1.51/1.36	.15/.11	<.001	1.68/1.91
Grammatical well-formedness	3.68/3.68	.17/.14	<.001	3.14/4.08

The effect sizes showed that Prosody and Plausibility effects were small/medium while all other effects were large. In order to assess the relative importance of the linguistic factors according to language, a second set of mixed 2 x 5 ANOVAs was carried out to analyse the mean differences between the Linguistics factors (5 levels: prosody, plausibility, lexical status of CWs/FWs, lexicality and grammaticality) and Language (2 levels: English and Czech).

The ANOVA revealed no main effect of Language ($F(1, 98) = 1.02$, $p = .314$, $\eta^2 = .01$), but a highly significant main effect of Linguistic factor ($F(3.47, 340.21) = 168.92$, $p < .001$, $\eta^2 = .63$), and no interaction between Language*Linguistic factor ($F(3.47, 340.21) = .82$, $p = .498$, $\eta^2 = .008$). Figure 2 provides an overview of results from post-hoc comparisons, with Bonferroni correction for multiple comparisons applied. As can be seen, Grammaticality had the most impact and differed significantly from all other factors (all ps <.001). Lexicality had a larger effect than Prosody ($p <.001$) and Plausibility ($p =.003$), but did not differ significantly from the CW/FW lexicality effect ($p = .519$). The CW/FW lexicality effect was significantly larger than Prosody ($p <.001$), but did not differ from Plausibility ($p = .848$). Plausibility and Prosody significantly differed from each other ($p = .018$).

Figure 2. Post-hoc comparisons of mean differences between linguistic factors according to language
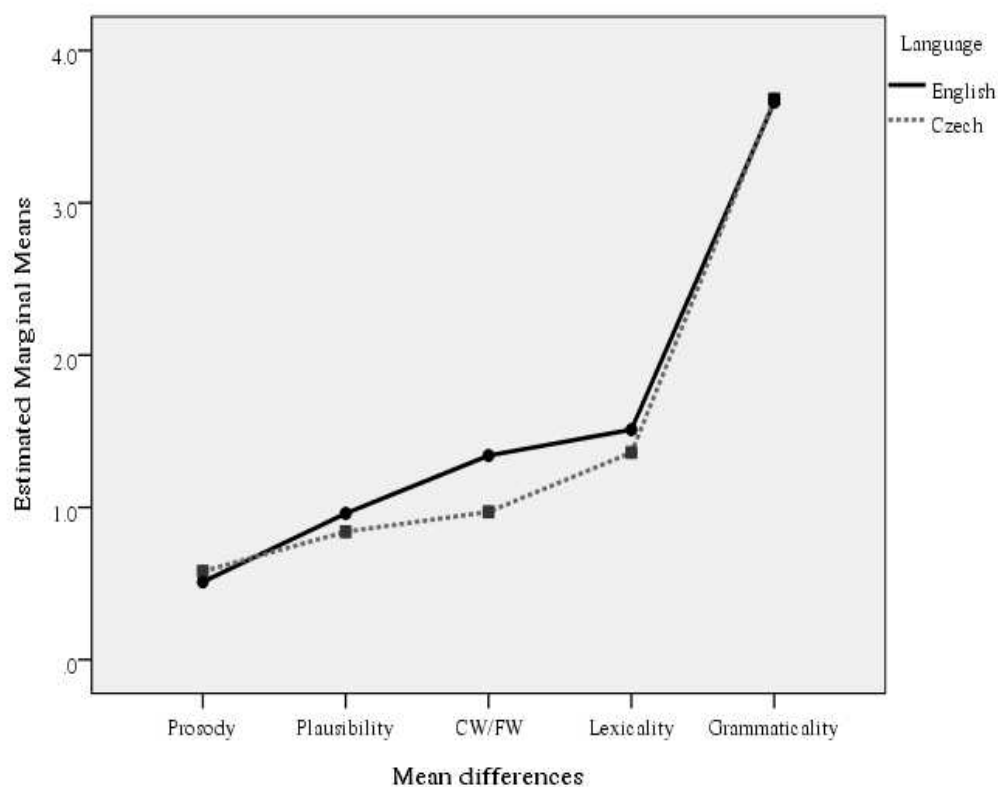
Figure 2 suggests that in English, real function words provide a larger benefit compared to the effect produced by plausibility, but this was not significant. This tendency could be because English FWs provide morphosyntactic frames for a sentence, while in Czech much of this information is expressed by inflectional morphology attached to CWs and therefore FWs do not play the same role.

3.3. Duration

Stimuli were carefully matched across the conditions for number of words and syllables, but duration of stimuli (in milliseconds) was not controlled. This raises the possibility that differences in duration might account for the differences in span. While we cannot deal with this issue in depth in this paper, we present some evidence that duration was not a key factor. Measurement of duration using Praat software revealed that many differences between conditions were matched by durational differences between same-length stimuli in those conditions. However, this was not always the case. Most notably, sentences with non-CWs vs non-FWs did not differ in duration but yielded mean spans differing by at least one word. Furthermore, since children attained higher spans in the easier conditions, the targets they repeated in these conditions were not only longer (in terms of number of words), but often similar or greater in duration than the targets achieved in more challenging conditions. For instance, while children's successful recall for ungrammatical sentences in English stopped at 4.35 words, with a mean duration of 1.88 seconds, their span for grammatical sentences with list prosody went up to 7.50 words, with a mean duration of 4.32 seconds. Together these results show that differences in duration of stimuli cannot account in any simple way for the differences we found across linguistic conditions.

4. Discussion

This study investigated the impact of different types of linguistic information on recall span in 4- to 5-year-old English- and Czech-speaking children. Our purpose was to clarify the nature of sentence repetition tasks and thereby inform their use in clinical assessment and increase our understanding of poor performance found in children with language impairment. For the first time, span was used as an outcome measure for recall at the sentence level. This allowed us to assess capacity for specific types of linguistic information found on a sentence level. The experiment yielded high levels of compliance and robust results across two typologically different languages.

4.1. The size of the contribution of linguistic factors to immediate recall

The results of this study throw new light on the interface between STM and long-term language knowledge. Our findings on the relative contribution of each type of linguistic domain reveal relations between STM and language that are established by 4-5 years and this lays the foundations for the theoretical modelling of their interface. Note that we are focusing on immediate verbatim recall, i.e. preservation of the just-heard phonological form of the lexical items and their linear sequence. In line with much previous evidence on children and adults, our findings demonstrate that verbatim recall in a familiar language is more than reproduction of a phonological string, since capacity varies according to the familiarity/knowledge of the material to be recalled. Every type of language knowledge was found to make a distinct and significant impact on immediate recall capacity, but some types of familiarity made a greater impact than others suggesting that they have a 'privileged' role in immediate recall. Unsurprisingly, and in line with previous research in adults and children (Miller & Isard, 1963; Scholes, 1970), recall increased when sentences were well-formed. The scale of the advantage is nonetheless striking (with effect size $d = 3.14$ for English and $d = 4.08$ for Czech), and demonstrates the extent to which morphosyntactic knowledge aids

sentence recall. Grammaticality (where function and content words are in expected positions, creating interpretable relations between these) had a larger effect than any other linguistic factor.

Our further finding that familiarity of function word frames was of more benefit than familiarity of content words adds to existing evidence on the importance of morphosyntax. It seems that familiar sequences of FWs facilitate segmentation and preservation of content word (or pseudocontent word) phonology for verbatim recall, at least in syntactically simple sentences like those presented in this study.

In contrast, plausibility had a significant but smaller effect (see Table 4), increasing span by less than one item in each language. This echoes Miller and Isard's (1963) finding that plausibility had substantially less effect than grammaticality on adults' learning of strings of words. Semantically plausible targets are conceptually more familiar than their implausible counterparts, and presumably contain sequences of items with a higher probability of co-occurrence, but these factors had limited benefit for verbatim recall.

The prosody effect was also significant and did not differ in size from the plausibility effect. Prosody may help in chunking and segmenting real word sequences and hence benefit recall capacity. We might presume, though, that prosodic chunking is largely redundant by 4 years, when the majority of children have mastered basic sentence structures in their language. This is consistent with findings from verbal learning in adults. O'Connell, Turner, and Onuska (1968) found that presenting stimuli with prosody was beneficial for adult recall, but the benefit was much smaller than the gains from the presence of morphosyntactic well-formedness. The picture may be different at an earlier age, when functions words are still emerging and children are likely to omit FWs in spontaneous production as well as repetition tasks (Scholes, 1970; Seeff-Gabriel et al., 2010). This possibility is supported by evidence that prosody affects function word repetition in 2-year-old children (Gerken, Landau, &

Remez, 1990).

Collectively, our findings demonstrate that immediate verbal recall is more reliant on familiarity of morphosyntax (well-formed sentence frames) than familiarity of lexical items and their semantics. We infer that immediate sentence repetition tasks are more a test of lexical phonology and morphosyntax than semantic/conceptual understanding.

4.2. Cross-linguistic comparisons

Our findings are strengthened by our cross-linguistic evidence. We did not conduct quantitative comparisons of absolute span scores due to the many differences between the Czech/English stimuli, including durational differences, number of morphemes, number of syllables, lexical properties and sample-related variables, which would confound interpretation of any differences in absolute levels of performance. Our comparison was between patterns of performance in the two languages, and these were almost identical. Grammaticality yielded the largest effect in both languages. In contrast, the effects of prosody and plausibility were small and did not differ from each other in either language. While the mean differences for Plausibility, Lexical status CW/FW and Lexicality were larger in English, the interaction between Language and Linguistic factor was not significant. Given the typological differences between English and Czech, particularly with respect to morphosyntactic structure and the role of FWs, parallels in our findings on the relative contribution of different types of language knowledge suggest these may reflect relatively universal aspects of language processing.

The purpose of the vocabulary task was to ensure that the children had language skills in the normal range. As shown in Table 1, the vocabulary raw scores in the younger age group were strikingly similar for English- and Czech-speaking children, but this was not the case in the older group; there was a tendency for Czech-speaking 5-year-old children to gain a higher

vocabulary score (p =.07). The difference might be related to the nature of the items and their lexical properties. While conceptually the items were similar, the information about lexical properties such as frequency or age of acquisition was not available for Czech and therefore not controlled in the Czech version. We might expect this to be less of an issue with the younger children as earlier items in the test are highly frequent and early acquired in both languages, and this appeared to be the case given the lack of a significant difference in the group of 4-year-olds. However, while the English items decrease in frequency and familiarity as the test progresses, it was often not possible to match this in the Czech version. For instance, 'feline' can only be translated as an adjective with the same lexical root as 'cat' which would be much more frequent and familiar and therefore Czech children would earn a point. It seems that the difference between the Czech and English 5-year-olds vocabulary skills at least partially reflected the differences in the tests.

4.3. Wider theoretical implications

Any theoretical model addressing the relationship between STM and LTM language knowledge needs to account for the range of linguistic structures that play a role in immediate recall, as recognized in the revised version of the working memory model by Baddeley (2000). Baddeley and colleagues acknowledge that the 'episodic buffer', added to their earlier model of working memory to accommodate the contribution of LTM, is underspecified (Allen & Baddeley, 2009; Baddeley, Hitch & Allen, 2009). Our findings contribute to further specification by providing evidence of the effects of different types of LTM language knowledge on immediate recall. This raises further issues for the Baddeley model, in particular: i) how the information from LTM might be transferred into this buffer, and ii) what kind of code is used for the storage of LTM material. Recent theories arguing for the intertwining of linguistic knowledge and memory in immediate recall tasks (Acheson &

MacDonald, 2009) might provide a theoretical framework for addressing this question. These theories invite investigation of how language processing and memory are intertwined, the issue at the heart of this study. By using span as a measure of recall, our findings have demonstrated the extent to which STM capacity increases with recognition of the linguistic material to be recalled, and suggest that immediate sentence recall engages 'surface' or formal aspects of sentence processing (lexical phonology and morphosyntax). In contrast, there is reason to think that delayed recall may rely more on semantics and pragmatics, which are crucial in comprehension and spontaneous production (Polišenská, Chiat, Comer & McKenzie, under review).

Our findings also bear on debates about the nature of SLI, in particular, the polarisation between 'limited capacity' theories which attribute SLI to a deficit in the amount of verbal information children are able to recall, and linguistic theories which attribute SLI to deficits in children's linguistic representations. Deficits in repetition tasks, which have been widely observed in children with SLI and are widely accepted as a clinical marker for SLI (Conti-Ramsden et al., 2001), have been interpreted as evidence for both positions: limited storage accounts and for limitations in linguistic representations. In demonstrating that verbal recall is inseparable from language processing and knowledge, our findings call into question this polarisation. Difficulties with sentence recall surely point to limitations in children's capacity for processing verbal information, but since this capacity is inseparable from representations arising from previous processing of linguistic input, the polarisation between capacity limitations and linguistic representations as the source of SLI becomes spurious. A similar conclusion has emerged from studies of relations between linguistic/metalinguistic processing and working memory as measured by the Competing Language Processing Task, a complex span task in which children are presented with a series of sentences to judge as true or false and their span for recall of the final word in each sentence is measured (Mainela-

Arnold et al., 2010; Mainela-Arnold et al., 2012). According to Mainela-Arnold and colleagues, 'hypotheses that causally link 'working memory capacity' to language ability are circular because the measure of working memory capacity (sentence span) is itself partially determined by language ability' (Mainela-Arnold et al., 2012, p.166). This conclusion shifts the focus for research from determining whether memory or language are the underlying problem in SLI, to the ways in which language processing and resulting linguistic representations impact on immediate and delayed recall in SLI.

4.4. Implications for clinical assessment

A key purpose of our research was to inform the use of immediate sentence recall in the assessment of children's language. Our findings strengthen the evidence base for the use of sentence repetition tests. The wide discrepancy between span for random words and well-formed sentences confirms that sentence repetition is much more than a test of STM: it draws on linguistic knowledge that children have acquired and provides evidence of their knowledge. Poor performance is therefore indicative of deficits in the domains of language that are most critical for immediate recall. It follows from our findings that sentence repetition should reveal deficits in lexical and morphosyntactic domains, but might be less sensitive to semantic/pragmatic deficits. It is striking that the effects of different conditions on spans did not change between 4 and 5 years in either language, although average spans increased over this age range. The age sensitivity of the task highlights its clinical potential in identifying delay; the consistency of patterns of performance, on the other hand, indicates its potential for identifying atypical processes in recalling sentences across age.

As an assessment tool, sentence repetition is inexpensive, quick and easy to administer, and has particular advantages for language communities that lack traditional language tests. Sentence repetition tasks can be easily adapted for different languages,

selecting stimuli to target key morphosyntactic properties of the language. Research has recently turned to the use of sentence repetition for assessment in bilingual populations (Chiat, Armon-Lotem, Marinis, Polišenská, Roy, & Seeff-Gabriel, 2013; Thordardottir & Brandeker, 2013). If immediate sentence repetition is most useful as an assessment of morphosyntactic skills, materials should be designed to minimise the possible impact of limitations in lexical knowledge due to limited exposure. For example, CWs should be selected to be maximally familiar, and should be phonologically simple to minimise effects of language-specific complexities in lexical phonology.

      In using immediate sentence recall for clinical assessment, it is important to understand what it is testing. Our findings identified linguistic skills that play the most significant role in immediate repetition, hence revealing the potential of this task in identifying language impairment in children, particularly deficits in lexical phonology and morphosyntax. An important next step is to investigate profiles of performance in children with language impairment which will throw more light on the sources of their difficulties with sentence repetition.

References

ACHESON, D. J., and MACDONALD, M. C., 2009, Verbal working memory and language production: Common approaches to the serial ordering of verbal information. Psychological Bulletin, **135**, 50-68.

ALLEN, R. J. and BADDELEY, A. D., 2009, Working memory and sentence recall. In A. Thorn and M. Page (eds), Interactions between short-term and long-term memory in the verbal domain (Hove: Psychology Press), pp. 63-85.

BADDELEY, A. D., 2000, The episodic buffer: A new component of working memory? Trends in Cognitive Sciences, **4,** 417-423.

BADDELEY, A. D., HITCH, G. J., and ALLEN, R. J., 2009, Working memory and binding in sentence recall. Journal of Memory and Language, **61**, 438–456.

BONVILLIAN, J. D., RAEBURN, V. P., and HORAN, E. A., 1979, Talking to children: the effects of rate, intonation, and length on children's sentence imitation. Journal of Child Language, **6**, 459–467.

CHIAT, S., ARMON-LOTEM, S., MARINIS, T., POLIŠENSKÁ, K., ROY, P., and SEEFF-GABRIEL, B., 2013, The potential of sentence imitation tasks for assessment of language abilities in sequential bilingual children. In V. C. Mueller Gathercole (ed), Bilinguals and assessment: State of the art guide to issues and solutions from around the world (Clevedon: Multilingual Matters), pp. 56-89.

CICCHETTI, D. V., 1994, Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, **6**, 284–290.

COADY, J. A., and EVANS, J. L., 2008, Uses and interpretations of non-word repetition

    tasks in children with and without specific language impairments (SLI). International

    Journal of Language and Communication Disorders, **43**, 1–40.

COHEN, J., 1988. Statistical power analysis for the behavioral sciences (Hillsdale: Erlbaum).

COLTHEART, M., 1981, The MRC psycholinguistic database. The Quarterly Journal of

    Experimental Psychology A: Human Experimental Psychology, **33A,** 497-505.

CONTI-RAMSDEN, G., BOTTING, N., and FARAGHER, B., 2001, Psycholinguistic

    markers for Specific Language Impairment (SLI). Journal of Child Psychology and

    Psychiatry, and Allied Disciplines, **42**, 741–748.

DEVESCOVI, A., and CASELLI, C. M., 2007, Sentence repetition as a measure of early

    grammatical development in Italian. International Journal of Language and

    Communication Disorders, **42,** 187–208.

DUNN, L. M., DUNN, L. M., WHETTON, C., and BURLEY, J., 1997, British picture

    vocabulary scale (2nd ed.) (BPVS-II) (Windsor, Berks: NFER-Nelson).

EVERITT, A., HANNAFORD, P., and CONTI-RAMSDEN, G., 2013, Markers for persistent

    specific expressive language delay in 3-4-year-olds: Markers for persistent specific

    expressive language delay in 3-4-year-olds. International Journal of Language and

    Communication Disorders, **48**, 534–553.

GATHERCOLE, S. E., FRANKISH, C. R., PICKERING, S. J., and PEAKER, S., 1999,

    Phonotactic influences on short-term memory. Journal of Experimental Psychology:

    Learning, Memory, and Cognition, **25,** 84–95.

GERKEN, L.A., LANDAU, B., and REMEZ, R. E., 1990, Function morphemes in young

    children's speech perception and production. Developmental Psychology, **26**, 204–216.

HALLGREN, K. A., 2012, Computing inter-rater reliability for observational data: An

    overview and tutorial. Tutorials in Quantitative Methods for Psychology, **8,** 23-34.

KIDD, E., BRANDT, S., LIEVEN, E., and TOMASELLO, M., 2007, Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. Language and Cognitive Processes, **22**, 860–897.

LEONARD, L. B., ELLIS, W. S., MILLER, C. A., FRANCIS, D. J., TOMBLIN, J. B., and KAIL, R.V., 2007, Speed of processing, working memory, and language impairment in children. Journal of Speech, Language, and Hearing Research, **50**, 408-428.

MAINELA-ARNOLD, E., and EVANS, J. L., 2005, Beyond capacity limitations: Determinants of word recall performance on verbal working memory span tasks in children with SLI. Journal of Speech, Language, and Hearing Research, **48**, 897–909.

MAINELA-ARNOLD, E., EVANS, J. L., and COADY, J., 2010, Beyond capacity limitations II: Effects of lexical processes on word recall in verbal working memory tasks in children with and without Specific Language Impairment. Journal of Speech, Language, and Hearing Research, **53**, 1656-1672.

MAINELA-ARNOLD, E., MISRA, M., MILLER, C., POLL, G. H., and PARK, J. S., 2012, Investigating sentence processing and language segmentation in explaining children's performance on a sentence-span task: Sentence span and language abilities. International Journal of Language & Communication Disorders, **47**, 166–175.

MARSHALL, C. R. and VAN DER LELY H, K. J., 2009, Effects of word position and stress on onset cluster production: Evidence from typical development, Specific Language Impairment, and Dyslexia. Language, **85,** 39–57.

MCDADE, H. L., SIMPSON, M. A., and LAMB, D. E., 1982, The use of elicited imitation as a measure of expressive grammar: a question of validity. The Journal of Speech and Hearing Disorders, **47,** 19–24.

METSALA, J. L., and CHISHOLM, G. M., 2010, The influence of lexical status and neighborhood density on children's nonword repetition. Applied Psycholinguistics, **31,** 489–506.

MILLER, G. A., and ISARD, S., 1963, Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior, **2**, 217–228.

O'CONNELL, D., TURNER, E., and ONUSKA, L., 1968, Intonation, grammatical structure, and contextual association in immediate recall. Journal of Verbal Learning and Verbal Behavior, **7,** 110–116.

PICKERING, S.J., and GATHERCOLE, S.E., 2001, Working Memory Test Battery for Children (London: Psychological Corporation Europe).

POLIŠENSKÁ, K., CHIAT, S, COMER, A., and MCKENZIE, K., under review, Semantic effects in sentence recall: The contribution of immediate vs. delayed recall in language assessment. Journal of Communication Disorders.

POLL, G. H., BETZ, S. K., and MILLER, C. A., 2010, Identification of clinical markers of specific language impairment in adults. Journal of Speech, Language, and Hearing Research, **53**, 414–429.

RICHES, N. G., 2012, Sentence repetition in children with Specific Language Impairment: an investigation of underlying mechanisms. International Journal of Language and Communication Disorders, **47**, 499–510.

ROY, P., and CHIAT, S., 2004, A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: Evidence from typically developing children. Journal of Speech, Language, and Hearing Research, **47**, 223–234.

SCHOLES, R., 1970, On functors and contentives in children's imitations of word strings. Journal of Verbal Learning and Verbal Behavior, **9**, 167–170.

SEEFF-GABRIEL, B., CHIAT, S., and DODD, B., 2010, Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. International Journal of Language and Communication Disorders, **45**, 691–702.

STOKES, S. F., WONG, A. M.-Y., FLETCHER, P., and LEONARD, L. B., 2006, Nonword repetition and sentence repetition as clinical markers of Specific Language Impairment: The case of Cantonese. Journal of Speech, Language, and Hearing Research, **49**, 219–236.

THORDARDOTTIR, E., and BRANDEKER, M., 2013, The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. Journal of Communication Disorders, **46**, 1–16.

WALKER, I., and HULME, C., 1999, Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. Journal of Experimental Psychology: Learning, Memory, and Cognition, **25**, 1256–1271.