The potential of sentence imitation tasks for assessment of language abilities in sequential bilingual children

Shula Chiat, Sharon Armon-Lotem, Theo Marinis, Kamila Polišenská, Penny Roy, & Belinda Seeff-Gabriel

Abstract

Sentence repetition tasks are increasingly recognised as a useful clinical tool for diagnosing language impairment in children. They are quick to administer, can be carefully targeted to elicit specific sentence structures, and are particularly informative about children's lexical and morphosyntactic knowledge. This chapter exlores the theoretical potential of sentence repetition for assessment of sequential bilingual children, and presents three studies comparing performance of sequential bilingual children with monolingual children's performance on standardised sentence repetition tests in Hebrew (children with L1 Russian, age 5-7 years, and L1 English, age 4½-6½ years), German (children with L1 Russian, age 4-7 years) and English (children with L1 Turkish, age 6-9 years). Results differed across studies: distribution of children in the Hebrew studies was in line with monolingual norms, while the majority of children in the English-Turkish study scored in a range that would be deemed impaired for monolingual children, and performance in the German-Russian study fell between these extremes. Analyses of performance within studies revealed similar discrepancies in effects of children's exposure to L2, with significant effects of Age of Onset in the Hebrew-Russian and Hebrew-English groups and some indication of Length of Exposure effects, but no effects of either factor in the English-Turkish group. Multiple differences between these studies preclude direct inferences about the reasons for these different results: studies differed in content, methods and scoring of sentence repetition tests, and in ages, languages, language exposure, and socioeconomic status of participants. It is possible that socioeconomic differences are associated with differences in language experience that are equally or more important than onset and length of exposure. Collectively, these studies demonstrate that sentence repetition provides a measure of children's proficiency in their L2, but that the use of sentence repetition in clinical assessment requires caution unless norms are available for the child's bilingual community. As a next step, it is proposed

that sentence repetition tests using early-acquired vocabulary and targeting aspects of sentence structure known to be difficult for monolingual children with language impairments should be developed in different target languages. This will allow us to explore further the factors that influence attainment of basic morphosyntax in sequential bilingual children, and the point at which sentence repetition, as a measure of morphosyntax, can help to identify children requiring clinical intervention.

The challenges of assessing language and identifying language impairment in bilingual children are all too familiar to clinicians. How can the child's language be meaningfully assessed if there are no standardised assessments of the child's L1, and performance on L2 assessments may reflect limited exposure to L2 rather than a developmental deficit? In this chapter, we make a case for the potential contribution of verbal imitation tasks – and more specifically sentence imitation tasks – in addressing these challenges. In the first part of the chapter, we review evidence that verbal imitation is highly informative about children's expressive language and has been found to differentiate children with SLI from typically developing children in diverse and typologically different languages. We then consider the role of language knowledge and experience in repetition tasks based on studies revealing the influence of linguistic and environmental factors on children's performance. The second part of the chapter presents studies of sentence imitation in four groups of children with different L1-L2 combinations. Based on theoretical and empirical evidence presented in the chapter, we put forward a multilingual agenda for developing and evaluating sentence imitation tasks to assess core language abilities in children's L2.

Verbal imitation as evidence of language abilities and deficits

In the early days of child language research, children's ability to imitate sentences attracted cursory attention, reflecting the theoretical perspective of the day. It was observed that children imitated 'within their system', be it phonological (Smith, 1973) or syntactic (Brown, 1973), but this served merely to supplement the rationalist argument that language could not be acquired by imitation of input, with empirical evidence that imitating the input does not help children achieve the adult model or attain the adult system. A few early investigations explored relations between imitation, comprehension and production of language in naturalistic and experimental situations (e.g. Fraser, Bellugi & Brown, 1963), but pursuit of imitation

in child language research and the impact on clinical assessment were limited. Exceptionally, the Carrow Elicited Language Inventory (Carrow, 1974) assesses sentence imitation in its own right, but there is little evidence of its use in clinical assessment or research.

When interest in children's ability to imitate surfaced again, some two decades later, it was in the very different guise of the nonword repetition task, which has come to occupy a privileged position in research on Specific Language Impairment (SLI) and developmental dyslexia. The nonword repetition test was originally developed by Gathercole and Baddeley (1989) as a relatively pure measure of phonological short-term memory (STM). The rationale for using nonwords as stimuli was that this would minimise the contribution of children's knowledge to their performance. The finding that nonword repetition performance related to children's wider language and reading abilities could then be advanced as evidence for the crucial contribution of phonological memory to the development of these complex skills.

The broad findings on nonword repetition for monolingual children are not in doubt. They have been replicated in a raft of studies reporting significant differences between typically developing monolingual children and children with SLI in English (see Coady & Evans, 2008; Gathercole, 2006; and Graf Estes, Evans, & Else-Quest, 2007 for overviews). Group differences have been reported for children at different ages and in different languages including Swedish (Sahlén, Reuterskiöld-Wagner, Nettelbladt & Radeborg, 1999), Italian (Casalini et al., 2007), and Spanish (Girbau & Schwartz, 2007). Indeed, the consistency of findings has led to the proposal that nonword repetition may serve as a clinical marker for SLI. To date, only one study has gone against the tide of crosslinguistic evidence, reporting no difference between children with and without SLI in Cantonese (Stokes et al., 2006).

While not as prolifically researched as nonword repetition, sentence repetition has been gaining attention as another possible marker of SLI in monolingual children. Clinical assessments of language often include a sentence recall subtest, as in the case of the Clinical Evaluation of Language Fundamentals (CELF; Wiig, Secord, & Semel, 1992; Semel, Wiig, & Secord, 1994) and the Test of Language Development (TOLD; Newcomer & Hammill, 1997). Most studies of sentence repetition in English have employed the Sentence Recall subtest of the CELF, and have found significant differences between children with SLI and typically developing (TD) peers (Bishop et al., 2009; Conti-Ramsden, Botting, & Faragher, 2001). In a comparison of four candidate clinical markers, Conti-Ramsden et al. found that Sentence Recall achieved the best combination of sensitivity and specificity. Studies using other sets of sentence stimuli have found significant differences between typically developing (TD), SLI, and other clinical groups (Willis & Gathercole, 2001; Redmond, 2005; Redmond, Thompson, & Goldstein, 2011; Riches et al., 2010). Seeff-Gabriel, Chiat and Dodd (2010) report on a sentence repetition test designed for children with severe speech difficulties. This test also yielded significant differences between children with SLI and TD peers, and Seeff-Gabriel et al. demonstrated its capacity to identify intact morphosyntactic abilities in children with unintelligible speech whose expressive language defied assessment using other methods. In the case of sentence repetition, even the Cantonese data fall into line: Stokes et al. report significant differences between their groups of children with SLI and age-matched TD peers on a test of sentence repetition in Cantonese.

The rationale for exploring sentence repetition tasks as a method for assessing children in their L2

Repetition tasks have a number of advantages that make them particularly attractive for L2 assessment. They are quick and easy to administer; in contrast to most language elicitation tasks, linguistic targets are explicit and precisely specified;

and they yield clear quantitative and qualitative results. Since targets are known and consistent, we can readily compare levels of performance and patterns of errors across children, making it easy to track children's progress in relation to typical performance whether in L1 or L2. These strengths are clearly exemplified by sentence repetition tasks. If carefully constructed, a set of 20 to 40 sentences allows sampling of a rich and representative range of sentence structures which it would be difficult, if not impossible, to elicit through a sentence production task or collection of spontaneous language data. The Sentence Imitation Test (SIT; Seeff-Gabriel, Chiat & Roy, 2008), for example, contains a comprehensive range of simple sentence structures sampling a comprehensive range of function words in English. This provides information about the aspects of sentences that children find difficult, which is very useful for therapy planning.

Alongside these general advantages, repetition tasks have a special role to play in L2 assessment, to the extent that they can be shown to be less affected by exposure and experience, which are known to be limited in L2. In the next section, we consider the ways in which linguistic characteristics of the stimulus materials affect children's repetition performance, revealing the benefits of language knowledge and by implication language experience. We then consider the effects of environmental differences on sentence repetition as evidenced by the performance of children from socially disadvantaged backgrounds who may be at risk of language disadvantage in their L1.

The contribution of language-specific knowledge to sentence repetition

It is now widely accepted that even nonword repetition, which was originally proposed as a 'pure' measure of phonological short-term memory, is subject to the familiarity of targets (Gathercole, 2006). All nonword repetition tests contain items that are consistent with the phonetics of the target language, being made up of consonants and vowels that occur in that language. However, tests vary substantially

in word-likeness of their nonwords (and even within tests, individual items may be more or less word-like): they may be characterised by more or less typical prosodic structure, more or less frequent phonotactic sequences, and may or may not contain morphemes of the language. All these factors turn out to influence children's nonword repetition performance (see Gathercole, 2006), demonstrating the effects of phonetic, phonological and morphological knowledge on the perception, storage and production of novel phonological forms. Nevertheless, influences on nonword repetition are clearly circumscribed compared with repetition of sentences, which open the floodgates to potential sources of support from language-specific knowledge. Sentences contain real words comprising fixed phonological forms with specific meanings, and these are combined according to the syntax and morphosyntax of the language to convey meaningful relations. Given the effects of phonology and morphology on nonword repetition, we might expect sentence repetition to be even more affected by familiarity with structures of the language.

A recent study by Polišenská (2011, under review) set out to address this issue by investigating the extent to which different types of language knowledge influence children's short-term memory span for sequences of words. Fifty typically developing Czech children and 50 typically developing English children aged 4 to 6 years participated in this study. The children were presented with blocks of successively longer sequences of words in different linguistic conditions, in order to determine their maximum span in each condition. The different conditions systematically varied syntactic, semantic, lexical and prosodic properties, as illustrated by the following examples of four-item length:

A	Well-formed sentence	*I hurt my knee*
B	Well-formed sentence with list prosody	*I, hurt, my, knee*

C	Semantically implausible sentence	*I dug my tea*
D	Pseudosentence with all lexical items replaced by nonwords	/ɔɪ vɜt kaɪ ri/
E	Syntactically ill-formed pseudosentence with sentence prosody	*Hurt my I knee*
F	Pseudosentence with content words replaced by nonwords	I /vɜt/ *my* /ri/
G	Pseudosentence with function words replaced by nonwords	/ɔɪ/ *hurt* /kaɪ/ *knee*

Comparison of conditions revealed the effect of each factor on immediate repetition performance. Prosodic structure (condition A vs B) had a significant effect, increasing memory span from a mean of 7.5 to 8.01 words in English and 7 to 7.58 in Czech, with a mean difference of 0.51 and 0.58, respectively. Semantic plausibility (condition A vs C) had a similar effect, increasing memory span from a mean of 7.05 to 8.01 words in English and 6.74 to 7.58 in Czech, a mean difference of 0.96 and 0.84, respectively. Familiarity of syntax and lexical items, however, produced the most notable effects. Lexical familiarity (condition A vs D), which brings with it morphosyntactic relations, dramatically increased memory span from a mean of 2.84 to 8.01 words in English and 2.54 to 7.58 in Czech, a mean difference of 5.17 and 5.04, respectively. Effects of morphosyntactic structure alone (condition A vs E) were almost as dramatic, with memory span increasing from a mean of 4.35 to 8.01 words in English and 3.9 to 7.58 in Czech, a mean difference of 3.66 and 3.68, respectively. Although the freer word order of Czech reduced possibilities for word order violations in the syntactically ill-formed condition, the effects of such violations were similar. Furthermore, despite its more limited repertoire of function words and greater reliance on inflections (which were not manipulated in this study), Czech showed a similar advantage for sentences containing real function words combined with nonwords in content word slots (condition F), compared with sentences containing

real content words combined with nonwords in function word slots (condition G). The difference between spans in these two conditions was statistically significant in both languages: 1.34 words in English and 0.97 in Czech. It seems that the form and distribution of closed class words, once acquired, are extremely robust and support immediate recall of verbal material.

Polišenská's findings demonstrate that sentence repetition draws on all aspects of sentence knowledge, but most distinctively on knowledge of syntax and morphosyntax. The implication is that children's repetition of simple sentences with familiar vocabulary is most informative about children's syntactic and morphosyntactic knowledge: their familiarity with words and the morphosyntactic devices that mark relations between these. Studies of early sentence repetition support this conclusion. In an investigation of sentence repetition in Italian TD preschoolers, Devescovi and Caselli (2007) found that omission of articles, prepositions and modifiers decreased between 2;0 and 2;6, and after 3 years of age, 'omissions of free function words practically disappeared' (p.188). Grammatical morphemes are known to be a particular challenge for children with language impairment, and persisting difficulties with grammatical morphology are frequently seen as a hallmark of SLI (Leonard, 1998). If simple sentence repetition is an effective test of grammatical morphology, it is unsurprising that it is a strong candidate for identifying SLI (see above).

Effects of variation in language experience on sentence imitation in L1

While the effects of knowledge and therefore experience of a specific language on repetition are clear, the effects of variations in experience *within* a language have received relatively little attention. Indeed, a key motivation for using nonword repetition as a clinical indicator is the claim that it tests processing skills that are relatively immune to prior knowledge and experience and therefore less biased against children from minority or disadvantaged backgrounds (Campbell et al., 1997;

Dollaghan & Campbell, 1998; Roy & Chiat, 2004). Accordingly, Campbell et al. found no difference between groups of first-graders from majority (White) and minority (primarily African American) backgrounds on a nonword repetition test. Engel, Santos and Gathercole (2008) compared nonword repetition performance in 6- to 7-year-old Brazilian children attending public and private schools. The groups were distinguished in terms of care-giver education, professional status, and income. They were found to differ on traditional measures of receptive and expressive language, but not on nonword repetition. The standardisation sample for the Early Repetition Battery (ERB; Seeff-Gabriel, Chiat, & Roy, 2008) showed very limited effects of socioeconomic status (SES) on the two repetition tests that make up the Battery: the Preschool Repetition Test (PSRep) and Sentence Imitation Test (SIT). Most notably, no significant effect was found for the key measures on the SIT: the number of content words and number of function words repeated correctly.

Based on these findings and others' evidence on nonword repetition in older children, Roy and Chiat (forthcoming) hypothesised that measures of verbal repetition might offer a key to cracking the well-recognised problem of differentiating limitations in language due to language disorder from limitations due to disadvantage. It was predicted that the distribution of performance on the PSRep and SIT in children from low SES backgrounds would be normal, in contrast to the downward shift in their scores on standard language measures. This hypothesis was investigated in a study of 387 children aged 3½ to 5 years, 219 attending nurseries/schools in Barking and Dagenham, a London borough with a high level of socioeconomic disadvantage, and 168 attending nurseries/schools in mid-to-high SES areas. The results were initially surprising: contrary to previous findings, the low and mid-to-high SES groups differed significantly on these tests, and the low SES group showed a clear deviation from the expected distribution. In the case of the SIT, 26.9% of the low SES children's function word scores fell more than 1 SD below the

mean, including 4.1% below -2 SDs. The skewing was even more marked for 'whole sentence correct' scores, with 31.5% of the low SES children more than 1 SD below the mean, including 7.8% below -2 SDs. Further analysis of performance according to age uncovered possible reasons for this unexpected finding, with important implications for the role of experience in verbal repetition. When the data were analysed by 6-month age bands, a cross-sectional developmental trajectory emerged: by age 4½ to 5, the gap between the expected and observed distribution of performance in the low SES group had narrowed and in some cases closed. Function word scores below -1 SD were down to 8.2%, including only 1.4% below -2 SDs, though scores for 'whole sentence correct' still found 20.6% of children below -1 SD. In contrast to the changes observed in the low SES group, the distribution in the mid-to-high SES group was stable across ages, with no children scoring below -2 SDs on either of the SIT measures, and over 95% performing within the normal range on both.

But how does this evidence square with the evidence that repetition skills are free of SES effects? The subtle evidence of linguistic influences on repetition performance observed in the Polišenská study may supply the missing piece. If detailed knowledge of lexical phonology and morphosyntax contributes to repetition performance, we might expect performance to vary with levels of experience even in L1. Given the lack of effects of SES in previous studies, it seems that variation in exposure for the range of SES groups compared in these studies is not sufficient to affect the emergence of core phonological and morphosyntactic knowledge, and/or the threshold for attaining this knowledge is relatively low (supported by the ERB standardisation results). The reduced performance in the Barking and Dagenham group, however, together with the catch-up in their first year of school, strongly suggests that prior to school this group's language experience was unusually limited,

affecting the deployment of core skills and acquisition of core phonological and morphosyntactic knowledge.

Implications for the investigation of sentence repetition tasks as a method for assessing children in their L2

We started this chapter by arguing that sentence repetition is an efficient and informative method for assessing children's language ability, and is therefore worth exploring as a method for L2 assessment. We have now seen that sentence repetition draws on language knowledge, particularly lexical and morphosyntactic knowledge, and is affected by extreme differences in social experience. This finding might lead us to conclude that sentence repetition is not after all a particularly useful avenue to explore for assessment of L2 children. Since these children are known to have late and/or reduced exposure to the language, how do we know whether a shortfall in sentence repetition performance is due to limited experience or to disorder? Our in-depth consideration of influences on sentence repetition has brought us back to the paradox of L2 assessment: the more a test draws on language-specific skills, the better it differentiates children with and without impairment; but the more it draws on language-specific skills, the more it depends on language experience that may be lacking in L2 children. Accordingly, sentence repetition is more clinically informative than nonword repetition using word-like items, which is in turn more informative than nonword repetition using non-word-like items (Archibald & Gathercole, 2006), but this advantage is counterbalanced by greater dependence on linguistic knowledge and therefore experience. Given the unavoidable trade-off between the two ideals for L2 assessment – optimal differentiation of children with language impairment and minimal reliance on language experience – sentence repetition may still have a special role to play. The potential for comprehensive and controlled sampling of targets affords precise and informative measurement of performance; this in turn affords detailed comparison

between groups and investigation of the ways in which between-group factors (such as amount and nature of exposure to L2, linguistic characteristics of L1 and L2) affect performance.

The importance of test targets and scoring

The impact of experience on a sentence repetition test will vary according to the demands it makes on language-specific knowledge and according to the criteria for scoring responses. As we saw with Polišenská's study, repetition of simple sentences is particularly subject to knowledge of lexical phonology, word order, and more specifically, the form and position of function words. In line with this, studies reported by Devescovi and Caselli (2007) and Seeff-Gabriel et al. (2010) revealed that measures of function words are particularly informative about language abilities in typically developing children and children with SLI. In considering the potential of sentence repetition for L2 assessment and the experience required to achieve L1 levels of performance, it will therefore be important to consider test content and how this is scored. In the most widely used measures of sentence repetition in English, the Sentence Recall subtests of the preschool and school-age Clinical Evaluation of Language Fundamentals (CELF; Wiig et al., 1992; Semel et al., 1994), targets vary in syntactic structure and become progressively more complex, but syntax is not systematically manipulated (see section on Turkish-English study below for examples).

In contrast to the Test of Language Development (TOLD; Newcomer & Hammill, 1997), which scores whole responses as right or wrong, CELF uses a more discriminating measure, scoring each response according to the number of errors made. As is clear from research findings, this scoring method is sensitive to group differences. However, the unspecified nature of the targets and purely quantitative scoring provide no information about the source of a child's difficulties or the nature of their errors.

Tests that systematically manipulate stimuli and/or employ more qualitative scoring methods are more informative about group differences. Riches et al. (2010), for example, manipulated the syntactic structure of sentences involving long-distance dependencies and employed a sensitive measure for scoring that took account of distance between the target sentence and the response. Using this scoring method, quantitative and qualitative differences were found between a group of adolescents with autism spectrum disorders plus language difficulties and a group with SLI. The SIT (Seeff-Gabriel et al., 2008) was specifically designed to assess morphosyntax in preschool children. This consists of 27 simple sentences that increase in length and complexity. Responses are scored in terms of number of content (open class) words, function (closed class) words and inflections repeated correctly. Each of these broad morphosyntactic classes can be broken down further into specific syntactic categories such as prepositions, auxiliary verbs, and pronouns.

The Cantonese sentence repetition test developed by Stokes et al. (2006) targeted two morphosyntactic structures, aspect and passive, and employed four scoring methods. Three were purely quantitative, while one scored 'core elements' of the target structures. Following findings on function words in English and Italian, we might expect the 'core element' score to be most discriminating and informative. Instead, Stokes et al. found that the CELF method of scoring sentences according to number of errors per sentence was the most effective in distinguishing children with and without SLI. It is not immediately obvious why scoring ‘core elements’ in Cantonese was less discriminating. This may reflect prosodic, semantic, and/or syntactic characteristics of the 'core elements' in Cantonese (as compared with characteristics of grammatical markers in English and Italian). Further insights into patterns of sentence repetition performance in Cantonese await detailed consideration of morphosyntactic categories and relations and the forms that express

these, together with evidence of repetition performance on these in TD children and children with SLI.

The Cantonese data provide a salutary reminder that test content will reflect morphosyntactic characteristics of the languages tested, which can vary considerably, and these may have effects on profiles of performance in different groups (typically and atypically developing, L1 and L2) and implications for clinical diagnosis. In the case of L2 children, linguistic characteristics of the L1 may also affect profiles of repetition performance in the L2. Considering this complex constellation of potential influences, any generalisations about the use of sentence repetition in L2 assessment will require comparisons across different and typologically diverse L1-L2 combinations (Gathercole, 2010), as well as variations in children's exposure to each language.

In the following sections of this chapter, we present exploratory investigations of L2 performance on existing sentence repetition tests in four L1-L2 communities: Russian-Hebrew, English-Hebrew, Russian-German, and Turkish-English. We compare the children's performance with monolingual norms, and consider effects of age of onset (AoO) and length of exposure (LoE) on performance.

Russian-Hebrew and English-Hebrew: A study of sentence repetition in sequential bilingual preschool children using Sentence Repetition with Pictures from the Goralnik Diagnostic Test of Hebrew

This section illustrates how a sample of sequential bilingual preschool children performed in the sentence repetition subtest of the *Goralnik Screening Test for Hebrew* (Goralnik, 1995), which is widely used in clinical settings in Israel for distinguishing between children with and without SLI . The Goralnik sentence repetition subtest has been normed for monolingual Hebrew-speaking children aged 3 to 6 years from high and low SES backgrounds, and provides norms for each SES group as well as combined norms. But do these norms hold for sequential bilingual

children? And how is L2 children's performance on this sentence repetition task affected by AoO and LoE?

Description of the sentence repetition task

The Goralnik sentence repetition subtest consists of 5 complex sentences, each describing a different picture, so that sentence repetition is presented in a pictorial context. The experimental sentences vary in length and complexity. In terms of length, the sentences range from 4 to 7 words, and from 7 to 11 morphemes, excluding verb inflections. Examples (1) and (2) illustrate the shortest (4 words, 7 morphemes) and longest (6 words, 11 morphemes) sentences as measured in morphemes:

(1) *ha-yeled mitnadned ve-ha-yalda oxelet*

The-boy swings and-the-girl eats

'The boy swings and the girl eats'

(2) *ha-yeled paxad she-ha-balon yauf lo me-ha-yad*

The-boy feared that-the-ballon will-fly to-him of-the-hand

'The boy was afraid that the balloon would fly out of his hand'

In terms of clausal complexity, all five sentences have two clauses and vary in the types of complexity, including coordination, finite and non-finite sentential complements, relative clauses, and direct speech, with one sentence for each structure. In terms of content, all sentences make use of basic vocabulary for topics familiar to preschool children, e.g. everyday events/actions involving people, animals, toys, and vehicles, as illustrated by the above examples.

Methodology

Participants

Seventy-five TD sequential bilingual Russian-Hebrew children (40 female) and 35 TD sequential bilingual English-Hebrew children (20 female) participated in the study. The Russian-Hebrew children had a mean age of 5;10 (range: 4;10-6;11,

SD: 6 months), and the English-Hebrew children a mean age of 5;9 (range: 4;5-6;6; SD: 6 months).

All children came from the same (mid-to-high) SES background, as defined by the mother's educational level. Both groups of L2 children were from L1 communities in the central part of Israel and attended preschools with no more than 50% children speaking their L1. They were growing up in families in which the language spoken at home was either Russian or English, and children did not have any history of speech and/or language delay or impairment, based on parental and school report.

Information about the L2 children's Age of Onset (AoO), Length of Exposure (LoE), and quantity and quality of input was collected through a parental and child questionnaire. AoO and LoE data for the two samples are presented in Table 1.

Table 1: Age of onset (AoO) and length of exposure (LoE) in the Russian-Hebrew and English-Hebrew samples

	AoO in months	Breakdown of sample according to AoO in years			LoE in months	Breakdown of sample according to LoE in years		
	Mean (SD) Range	<2;0	2;0-3;0	>3;0	Mean (SD) Range	1-2	3-4	>4
Russian-Hebrew (n=75)	35 (18) 0-66	22	24	29	37 (18) 12-75	23	34	18
English-Hebrew (n=35)	35 (13) 6-51	7	6	22	34 (16) 15-69	14	16	5

Procedure

The children participated in a battery of standardised and non-standardised assessments and experimental tasks. Testing was carried out in a quiet room in the preschool. The sentences of the sentence repetition subtest from the Goralnik test were presented to the children by a native speaker of Hebrew. Children were told that they had to repeat each sentence verbatim. The children's responses were audio-recorded and were also scored manually on an answer sheet during the session.

Scoring

The children's responses were scored using the guidelines from the Goralnik manual, which awards a score of 6, 3 or 0 for each response. A score of 6 is given

for a verbatim repetition (e.g. *ha-yeled axal tapuax ve-ha-yalda kar'a sefer* 'The boy ate an apple and the girl read a book'), a score of 3 for repetition of all major components with minor deviations (e.g. *ha-yeled axal tapuax ve-hi kar'a sefer* 'The boy ate an apple and she read a book'), and a score of 0 for a repetition that lacks some of the major constructs, such as subject, verb, or object, that appear in the original sentence (e.g. *ha-yeled axal ve-ha-yalda kar'a sefer* 'The boy ate and the girl read a book'). The final score is the sum of the scores on all items (maximum=30). The Goralnik provides monolingual norms for raw scores. Z-scores were calculated based on the monolingual norms. The z-scores reflect the distance from the mean score in SDs.

Results

Table 2 shows the mean z-scores, standard deviations and ranges for the Russian-Hebrew and English-Hebrew children on the sentence repetition task. The mean scores for both groups were within the monolingual normal range.

Table 2: Results on Goralnik sentence repetition for Russian-Hebrew and English-Hebrew groups

	Russian-Hebrew *(N = 75)*			English-Hebrew *(N =35)*		
	Mean	SD	Range	Mean	SD	Range
z-score	0.11	(0.89)	-2.36 to 1.05	0.31	(0.75)	-1.69 to 1.28

Breakdown of Russian-Hebrew and English-Hebrew sentence repetition performance according to monolingual norms

To investigate how children performed in terms of monolingual norms, we calculated the number of children who scored above the monolingual normal range

(above 1 SD), within the monolingual normal range (below 1 SD and above -1 SD), between 1 and 2 SDs below the mean, and at or below -2 SD. Results are illustrated in Table 3.

Table 3: Number of children scoring above, within and below normal range

	Russian-Hebrew *(N = 75)*	English-Hebrew *(N =35)*
1 or more SD above	10 (13%)	4 (11%)
Within normal range	56 (75%)	28 (80%)
1 to 2 SD below	6 (8%)	3 (9%)
2 SD or more below	3 (4%)	0 (0%)

Most of the bilingual children performed within the normal range; only 12 children in the combined sample performed below -1 SD, of whom only 3 children scored below -2 SDs. The profile of performance in these groups was therefore largely in line with norms and in fact slightly skewed towards the upper end (which might reflect the nature of the sample, with both groups from mid-to-high SES backgrounds and excluding children with any history of difficulties).

Analysis of performance by AoO

To investigate effects of AoO, we split the groups of L2 children into one group with AoO between 0;0 and 2;0, a second group with AoO between 2;1 and 3;0, and a third group with AoO between 3;1 and 5;4, as shown in Table 4.

Table 4: Breakdown of sentence repetition performance by AoO

	AoO 0;0 to 2;0				AoO 2;1 to 3;0				AoO 3;1 to 5;4			
	N	Mean	SD	Range	N	Mean	SD	Range	N	Mean	SD	Range
Russian-Hebrew	22	0.51	(0.43)	-0.16 to 1.05	24	0.11	(0.79)	-1.75 to 1.05	29	-0.19	(1.11)	-2.36 to 1.05
English-Hebrew	7	0.88	(0.21)	0.39 to 1.01	6	0.44	(0.48)	-1.6 to 0.94	22	0.14	(0.82)	-1.69 to 1.28

For the Russian-Hebrew bilinguals, a one-way ANOVA showed a significant difference (F (2 ,72) = 4.212, p = 0.02) traced by a post hoc Tukey test to a significant difference between the group with the lowest AoO and that with the highest AoO (p < 0.02). Similarly, for the English-Hebrew bilinguals, a one-way ANOVA showed a significant difference (F (2 ,32) = 3.413, p < 0.05), traced by a post hoc Tukey test to a significant difference between the group with the lowest AoO and that with the highest AoO (p < 0.05).

To investigate effects of AoO on how children performed in terms of monolingual norms, we merged the two samples and then calculated the number of children who scored above the monolingual normal range (above 1 SD), within the monolingual normal range (below 1SD and above -1 SD), between 1 and 2 SDs below the mean, and at or below -2 SD within each AoO range (see Table 5).

Table 5: Number of children scoring above, within and below normal range according to AoO

	AoO 0;0 to 2;0 (N=29)	AoO 2;1 to 3;0 (N=30)	AoO 3;1 to 5;4 (N=51)
1 or more SD above	6(21%)	1 (3%)	7 (13%)
Within normal range	23 (79%)	26 (87%)	35 (69%)
1 to 2 SD below	0 (0%)	3 (10%)	6 (12%)
2 SD or more below	0 (0%)	0 (0%)	3 (6%)

All children with AoO of 2;0 and below performed within or above the monolingual normal range. Of the 30 children with AoO of 2;1 to 3;0, only 3 (10%) performed below the monolingual normal range; however, of the 51 children with AoO of 3;1 to 5;4, 9 (17.65%) performed below the monolingual normal range. Notably, all of those who performed at least 2 SD below the mean were in the latest AoO group.

Analysis of performance by LoE

To investigate effects of LoE, we split the groups of L2 children into three groups with 1 to 2 years of exposure, 2;1 to 4 years of exposure and 4;1 to 6 years of exposure, as shown in Table 6.

Table 6: Breakdown of sentence repetition performance by LoE

	LoE 1 to 2 years				LoE 2;1 to 4 years				LoE 4;1 to 6 years			
	N	Mean	SD	Range	N	Mean	SD	Range	N	Mean	SD	Range
Russian-Hebrew	23	-0.08	(1.07)	-2.36 to 1.05	34	0.04	(0.90)	-2.36 to 1.05	18	0.47	(0.48)	-0.16 to 1.05
English-Hebrew	14	-0.10	(0.79)	-1.69 to 1.28	16	0.52	(0.64)	-1.25 to 1.01	5	0.83	0.24	0.39 to 0.94

All children with more than 4 years of exposure were within or above the monolingual normal range, and the three children who performed more than 2 SD below the monolingual mean had less than 4 years of exposure. For the Russian-Hebrew bilinguals, a one-way ANOVA showed no significant difference between the performance of L2 children in the different exposure groups (F (2,72) = 2.137, p = 0.125). For the English-Hebrew bilinguals, a one-way ANOVA showed a significant difference between the performance of L2 children (F (2,32) = 4.878, $p < 0.02$), traced by a post hoc Tukey test to a significant difference between the shortest exposure group and the longest exposure group ($p < 0.05$), and between the shortest exposure group and the medium exposure group ($p < 0.05$), with no significant difference between the medium and longest exposure groups.

Summary of the Russian-Hebrew and English-Hebrew findings

The findings in the L2 Hebrew studies (with L1 Russian or English) show that the majority of children (88% and 91% respectively) perform within or above the normal range for monolingual children despite variations in exposure. Only three out of a combined group of 110 children, all three of whom were exposed to Hebrew after

the age of three, fell more than 2 SDs below the monolingual mean. The overall distribution of performance of this bilingual group in the L2 is therefore in line with the normal monolingual distribution, making sentence repetition a promising tool for use in L2 assessment of this population. Nevertheless, results indicate that age of onset and length of exposure need to be taken into account, and some caution exercised in interpreting low scores when a child's exposure to the language has been limited.

Russian-German: A study of sentence repetition in sequential bilingual preschool children using the Sprachstandscreening für das Vorschulalter (Grimm, 2003) for German

This section illustrates how a sample of sequential bilingual preschool children performed on the Sprachstandscreening für das Vorschulalter (Grimm, 2003), a sentence repetition task that is used in clinical settings in Germany (Berlin) for distinguishing between children with and without SLI. The sentence repetition task has been normed for monolingual children from high and low SES backgrounds. Again, the question is whether these norms hold for sequential bilingual children, and how the L2 children's performance on this sentence repetition task is affected by the AoO and LoE.

Description of the sentence repetition task

The Sprachstandscreening für das Vorschulalter consists of 15 sentences, six semantically sensible sentences and nine semantically anomalous sentences, which vary in length and complexity. In terms of length, the sentences range from 6 to 10 words. Examples (3) and (4) illustrate the shortest and longest sentences as measured in words:

(3) *Die Ente sitzt neben dem Auto*

the duck sits beside the car

'The duck is sitting beside the car'

(4) *Der Schmutzige Hund wird vom Vater in der Wanne gebadet*

the dirty dog is by the father in the bathtub bathed

'The dirty dog is being bathed by the father in the bathtub'

In terms of sentence complexity, of the six semantically sensible sentences, four sentences comprised simple clauses (two in the passive voice, as exemplified by (4) above), and two had adverbial modifiers (one of which was a whole clause):

(5) *Vor dem Schlafen putzen Kinder die Zähne*

Before the sleep brush children the teeth

'Before sleeping the children brush their teeth'

(6) *Die Kinder lachen, weil sie auf dem Bett hüpfen*

The children laugh because they on the bed jump

'The children are laughing because they are jumping on the bed'

Of the semantically anomalous sentences, five sentences comprised simple clauses (two in the passive voice), e.g. (7), and four sentences had two clauses, two with a relative clause and two with adverbials, e.g. (8):

(7) *Der Kindergarten wird von den roten Bären geschütt*elt

the kindergarten is by the red bears shaken

'The kindergarten is shaken by the red bears

(8) *Bevor der Goldfisch hinfällt, frisst er aus dem Fenster*

before the goldfish falls, eats it off the window

'Before the goldfish falls down, it eats from the window.'

In terms of content, all sentences made use of basic vocabulary for topics familiar to preschool children, e.g. everyday events/actions involving people, animals, furniture, vehicles, as illustrated in the above examples. Unlike the Hebrew standardized test, the sentences in the German standardized test are presented without pictures.

Methodology

Participants

Sixty-one typically developing (TD) sequential bilingual Russian-German children (30 female) participated in the study. These children had a mean age of 5;6 (range: 3;11-7;2, SD: 10 months). In terms of SES, calculated by parents' education, 25 out of the 61 fathers and 32 out of 61 mothers had less than 12 years of education (i.e. high school), while the other parents had more than 13 years of education (i.e. high school and some further education). This yielded two SES groups, high and low, as measured by mothers' educations.

The Russian-German children were from the Russian community in Berlin, Germany, and attended preschools with no more than 50% Russian-speaking children. They were growing up in families in which the language spoken at home was Russian, and had no history of speech and/or language delay or impairment based on parental and school report.

Information about the L2 children's Age of Onset (AoO), Length of Exposure (LoE), quantity and quality of input was collected through a parental and child questionnaire. AoO and LoE data for this sample are presented in Table 7.

Table 7: Age of onset (AoO) and length of exposure (LoE) in the Russian-German sample

	AoO in months	Breakdown of sample according to AoO in years			LoE in months	Breakdown of sample according to LoE in years		
	Mean (SD) Range	<2;0	2;0 to 3;0	>3;0	Mean (SD) Range	1 to 2	3 to 4	>4
Russian-German (n=61)	24 (10) 0-46	37	14	10	42 (15) 13-82	7	31	23

Procedure

The children participated in a battery of standardised and non-standardised assessments and experimental tasks. Testing was carried out in a quiet room in the preschool. The sentences of the German sentence repetition task were presented to the children by a native speaker of German, and they were told that they had to repeat each sentence verbatim. The children's responses were audio-recorded and were also marked manually on an answer sheet during the session.

Scoring

The children's responses were scored using the guidelines from the Sprachstandscreening für das Vorschulalter (Grimm, 2003) manual.

To calculate raw scores, the number of words repeated correctly was counted. Z-scores were calculated based on the monolingual norms. The z-scores reflect the distance from the mean monolingual score in SDs.

Results

Comparison between low SES and high SES Russian-German children

Table 8 shows the mean z-score, standard deviation and range for the Russian-German children on the sentence repetition task according to SES as measured by mothers' education.

Table 8: Results on German sentence repetition test according to SES

	Lower SES *(N = 32)*			Higher SES *(N =29)*		
	Mean	SD	Range	Mean	SD	Range
Z-score	-0.76	(0.66)	-2.20 to 0.4	-0.36	(1.02)	-2.90 to 1.60

A one-way ANOVA showed a near-significant difference between the two SES groups (F (1 ,59) = 3.445, p = 0.068). This becomes significant if the cut-off point is 14 years of education rather than 12 years (F (1,59) = 4.667, p = 0.035). Nevertheless, the mean score for both groups was within the monolingual normal range.

Breakdown of Russian-German sentence repetition performance by SES according to monolingual norms

To investigate how children performed relative to monolingual norms, we calculated the number of children who scored above the monolingual normal range (above 1 SD), within the monolingual normal range (below 1 SD and above -1 SD),

between 1 and 2 SDs below the mean, and at or below -2 SD. Results are illustrated in Table 9.

Table 9: Number of children scoring above, within and below normal range

	Lower SES *(N = 32)*	Higher SES *(N =29)*	Total *(N=61)*
1 SD above	0 (0%)	3 (10%)	3 (5%)
Within norms	19 (59%)	21 (73%)	40 (66%)
1 to 2 SD below	12 (37%)	3 (10%)	15 (24%)
2 SD or more below	1 (4%)	2 (7%)	3 (5%)

Two-thirds of the bilingual children performed within or above the monolingual normal range. While a third of the sample performed below -1 SD, only 3 children's performance was more than 2 SDs below the mean. Notably, though, in the lower SES group only 56% of the children performed within the monolingual normal range, compared with 76% of the children in the higher SES group.

Analysis of performance by AoO

To investigate effects of AoO, we split the group of L2 children into one group with AoO between 0;0 and 2;0 (n=37), a second group with AoO between 2;1 and 3;0 (n=14), and a third group with AoO between 3;1 and 3;10 (n=10) as shown in Table 10.

Table 10: Breakdown of sentence repetition by AoO

AoO 0;0 -2;0 *(N=37)*			AoO 2;1-3;0 *(N=14)*			AoO 3;1-3;10 *(N=10)*		
Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
-0.51	(0.81)	-1.70 to 1.60	-0.68	(0.96)	-2.90 to 0.90	-0.65	(1.01)	-2.30 to 0.40

For the Russian-German bilinguals, a one-way ANOVA showed no significant difference for AoO (F (2 ,58) = 0.241, p = 0.78). Introducing the mother's education as a covariate had no impact on the results.

To investigate effects of AoO on how children performed in terms of monolingual norms, we calculated the number of children who scored above the monolingual normal range (above 1 SD), within the monolingual normal range (below 1SD and above -1 SD), between 1 and 2 SDs below the mean, and at or below -2 SD within each AoO range (see Table 11).

Table 11: Number of children scoring above, within and below normal range according to AoO

	AoO 0;0-2;0 *(N=37)*	AoO 2;1-3;0 *(N=14)*	AoO 3;1-3;10 *(N=10)*
1 SD above	3 (8%)	0(0%)	0 (0%)
Within norms	24 (65%)	9 (64%)	7 (70%)
1 to 2 SD below	10 (27%)	4 (29%)	1(10%)
2 SD or more below	0 (0%)	1 (7%)	2 (20%)

Three of the children with AoO of 2;0 or below performed above the monolingual normal range, and all the others were either within the monolingual normal range or no more than 1.7 SD below the mean. Of the 14 children with AoO of 2;1 to 3;0, 5 performed below the monolingual normal range, and of the 10 children with AoO of 3;1 to 3;10, 3 performed below the monolingual normal range. Notably, all of those who performed 2 SD below the mean were in the lower AoO groups.

Analysis of performance by LoE

To investigate effects of LoE, we split the groups of L2 children into three groups with 1 to 2 years of exposure, 2;1 to 4 years of exposure, and 4;1 to 5;9 years of exposure, as shown in Table 12.

Table 12: Breakdown of sentence repetition by LoE

LoE 1;0 to 2;0 years *(N=7)*			LoE 2;1 to 4;0 years *(N=31)*			LoE 4;1 to 5;9 years *(N=23)*		
Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
-1.01	(0.77)	-2.30 to 0.10	-0.66	(0.92)	-2.90 to 1.60	-0.31	(0.76)	-1.70 to 1.20

Most of the children with more than 4 years of exposure (87%) were within or above the monolingual normal range, and the three children who performed more than 2SD below the monolingual mean had less than 4 years of exposure, two of them with less than 2 years of exposure. A one-way ANOVA showed no significant difference, however, for LoE (F (2,58) = 2.243, p = 0.115). Introducing the mother's education as a covariate yields a near-significant difference (F (3,57) = 2.434, p = 0.074).

Summary of Russian-German findings

The findings in the L2 German studies (with L1 Russian) show that overall two-thirds of the bilingual children performed within or above the monolingual normal range, while 5% scored more than 2 SD below the mean. This distribution is notably

lower than the normal monolingual distribution (especially considering the sample excluded children with language difficulties). The distribution in the higher SES group came closer to matching the normal monolingual distribution and included 3 children scoring above the normal range, but even so, higher than expected numbers fell below -2 SD (7%, compared with expected 2.3% in the normal monolingual distribution). Those scoring below -2 SD were exposed to German after the age of two and for less than four years. These outcomes demonstrate the need for some caution in using monolingual norms with this population, and the possible influence of SES, as well as language exposure, on children's performance.

Turkish-English: A study of sentence repetition in sequential bilingual children using the Sentence Recall subtest from CELF-3

This section illustrates how a sample of sequential Turkish-English bilingual children performed in the Sentence Recall subtest of the Clinical Evaluation of Language Fundamentals III (CELF-3) (Semel, Wiig & Secord, 2000). The CELF is widely used in clinical settings in the UK, and the Sentence Recall subtest has been shown to have a high degree of specificity and sensitivity in distinguishing between children with and without SLI (Conti-Ramsden et al., 2001). However, it is unclear whether or not this will hold for sequential bilingual (L2) children. In order to address this question, we need to establish whether typically developing L2 children score within monolingual norms, and how their performance relates to their age of onset (AoO) and length of exposure (LoE). The following study compared the performance of typically developing Turkish-English children on CELF Sentence Recall relative to monolingual children, and investigated the effects of AoO and LoE.

Description of CELF Sentence Recall

The CELF Sentence Recall subtest consists of two practice and 26 test sentences. The test sentences vary in length, complexity, and content. In terms of

length, the sentences range from 6 to 19 words. Examples (9) and (10) illustrate the shortest and longest sentence:

(9) Did the girl catch the netball?

(10) The boy [who didn't turn up for practice] wasn't allowed to play for the team until a week later.

In terms of clausal complexity, sentences range from one to five clauses. (9) exemplifies a sentence with one clause, while examples (10)-(13) illustrate complex sentences. Sentence (10) includes a subject relative clause, (11) involves sentence coordination, (12) a truncated passive structure in embedded and main clauses, and (13) five clauses including one double embedding:

(11) [The fielder caught the ball] and [the crowd cheered loudly]

(12) [Before the first years were dismissed for lunch] [they were told to hand in their assignments].

(13) [When the students had finished studying] [they decided [to get something [to eat]] [before going home]].

In terms of pragmatic and lexical content, 16 out of the 26 sentences had a school-related topic involving school-related vocabulary (as illustrated by the above examples), and the remaining sentences were on other topics, e.g. games, everyday events involving people, and events involving animals and vehicles.

It is important to note that the factors of length, complexity, and topic are confounded in this task. The first six sentences on the test are relatively short, consist of one clause, and most involve events that are not school-based. All other sentences are long, consist of multiple clauses, and their content almost always relates to activities around school. If children are successful in repeating the first six sentences and make errors in the remaining sentences, it is impossible to know whether they fail because of the length of the sentences, their complexity, or a lack of familiarity with the school-related vocabulary.

Methodology

Participants

Seventeen TD sequential bilingual Turkish-English children (8 female) and 15 TD age-matched monolingual English-speaking children (10 female) participated in the study. The L2 children had a mean age of 7;10 (range: 6;1-9;3; SD: 13 months), and the L1 children a mean age of 7;10 (range: 7;2-8;11; SD: 5 months) (F (1, 30) = 0.36, p = 0.851). Both groups of children attended schools whose percentage of free school meals was well above the national average, indicating low socioeconomic status. The L2 children were from the Turkish community in London, were growing up in families in which the language spoken at home was Turkish, and attended schools with a high density of Turkish-speaking children. The monolingual children were attending schools in Reading. Both samples excluded children with a history of speech and/or language delay or impairment based on parental and school report.

Information about the L2 children's age of onset (AoO), length of exposure (LoE), and quantity and quality of input was collected through a parental and child questionnaire. AoO and LoE data are presented in Table 13.

Table 13: Age of onset (AoO) and length of exposure (LoE) in the Turkish-English sample (n=17)

AoO in months	Breakdown of sample according to AoO in years		LoE in months	Breakdown of sample according to LoE in years	
Mean (SD) Range	<3;0	3;0-5;0	Mean (SD) Range	1-3	4-6
39 (6.6) 29-60	9	8	55 (17) 21-83	5	12

Procedure

The children participated in a battery of standardised and non-standardised assessments and experimental tasks. The Sentence Recall task from CELF-3 was recorded in the Speech Booth, a purpose-built sound-proof room, at the Department of Clinical Language Sciences, University of Reading. The sentences were digitally recorded at normal speed by a female speaker and merged into a single file using Adobe Audition.

A laptop was used for the presentation of this task. Sentences were presented to the children through headphones, and a microphone connected to the laptop was used to record the children's responses. This task does not include pictures. The children were told that they had to repeat each sentence verbatim. Their responses were recorded using Adobe Audition.

Scoring

The children's responses were scored as specified in the CELF manual. To calculate raw scores, each sentence is given a score of 3 if it is repeated verbatim, a score of 2 if there is one error, a score of 1 if there are two or three errors, and a score of 0 if there are four or more errors. Standard scores were calculated based on the monolingual norms from CELF (mean = 10, SD = 3). Z-scores were calculated from the standard scores.

Results

Comparison between L2 and L1 children

Table 14 reports the standard and z-scores of the L2 and L1 children on the Sentence Recall task.

Table 14: Results on CELF Sentence Recall for L1 and L2 groups

	L1 *(N = 15)*			L2 *(N =17)*		
	Mean	SD	Range	Mean	SD	Range
Standard score	9.5	(2.3)	5-13	4.5	(2.1)	3-10
Z-score	-0.2	(0.8)	-1.7 to 1	-1.8	(0.7)	-2.3 to 0

While the mean for the L1 group fell just short of the population mean, the L2 group mean was almost two standard deviations below. A one-way ANOVA using the standard scores confirmed that the L2 children were significantly less accurate than the L1 children ($F(1, 31) = 45.29$, $p < 0.001$).

Breakdown of L1 and L2 Sentence Recall performance according to monolingual norms

To compare the distribution of performance in each group with the monolingual normal distribution, we calculated the number of children that scored above the monolingual normal range (above 1 SD), within the monolingual normal

range (below 1 SD and above -1 SD), between 1 and 2 SDs below the mean, and at or below -2 SD. Results are illustrated in Table 15.

Table 15: Number of children scoring above, within and below normal range

	L1 *(N = 15)*	L2 *(N =17)*
1 SD above	0	0
Within norms	13 (87%)	2 (12%)
1 to 2 SD below	2 (13%)	3 (18%)
2 SD or more below	0 (0%)	12 (70%)

It is notable that no child in these samples scored above the monolingual normal range. However, most of the monolingual children (87%) performed within the L1 normal range, with only two children (13%) performing between 1 and 2 SDs below the L1 mean, and no child scoring below -2 SD. The exact opposite pattern was observed in the group of L2 children. Only two children (12%) performed within the L1 normal range, while 3 children (18%) performed between 1 and 2 SDs below the L1 mean, and the majority (70%) performed below -2 SDs.

Analysis of L2 performance by AoO and LoE

To investigate whether or not AoO affected the L2 children's performance, we divided the group of L2 children into groups with AoO between 2;5 and 3;0 years (n=9) and between 3;3 and 5 years (n=8). Results are shown in Table 16.

Table 16: Breakdown of L2 Sentence Recall performance according to AoO

	AoO 2;5-3;0 *(N =9)*			AoO 3;3-5;0 *(N =8)*		
	Mean	SD	Range	Mean	SD	Range
Standard score	5	(2.5)	3-10	4	(1.3)	3-7
Z-score	-1.7	(0.8)	-2.3 to 0	-2	(0.4)	-2.3 to -1

A one-way ANOVA using the standard scores showed no significant difference between the performance of L2 children with lower vs. higher AoO (F (1,16) = 1.54, p = 0.233).

To investigate whether LoE affected the L2 children's accuracy, we divided the children into a group with 1 to 3 years of exposure (n=5) and a group with 4 to 6 years of exposure (n=12). Results are shown in Table 17.

Table 17: Breakdown of L2 Sentence Recall performance according to LoE

	LoE 1 to 3 years *(N =5)*			LoE 4 to 6 years *(N =12)*		
	Mean	SD	Range	Mean	SD	Range
Standard score	4	(0)	4	4.8	(2.5)	3-10
Z-score	-2	(0)	-2	-1.7	(0.8)	-2.3 to 0

A one-way ANOVA using the standard scores showed no significant difference between the performance of L2 children with shorter vs longer exposure (F (1,16) = 0.69, p = 0.421).

Analysis by sentence type

As pointed out above, the sentences in the CELF Sentence Recall task become progressively longer and more complex. Given the sensitivity of sentence repetition to morphosyntactic knowledge in simple sentences (see above), it is useful

to investigate how length/complexity affected performance in the L1 and L2 groups. Table 18 shows the scores for the short/simple sentences (1-6) vs. the long/complex sentences (7-26) in the two groups in numbers (maximum score of short sentences: 18; maximum number of long sentences: 60) and in percentage correct.

Table18: Mean raw scores and percentages on short/simple vs. long/complex sentences according to group

		L1 *(N = 15)*			L2 *(N =17)*		
		Mean	SD	Range	Mean	SD	Range
Short/simple sentences (max=18)	N	17.5	(1.3)	13-18	14.6	(3.2)	9-18
	%	97%	(7.2)	72-100%	81.4%	(17.6)	50-100%
Long/complex sentences (max=60)	N	18.1	(7.7)	6-33	4.2	(6.5)	0-21
	%	30.2%	(12.8)	10-55%	7.1%	(10.8)	0-35%

To compare the groups on the two sentence types, we conducted a mixed ANOVA on the percentage correct with Group (L2 vs. L1) as a between-subjects factor and Sentence Type (short/simple vs. long/complex) as a within-subjects factor. This showed a main effect of Group (F (1,30) = 24.07, $p < 0.001$), reflecting lower accuracy of the L2 children, and a main effect of Sentence Type (F (1,30) = 1021.26, $p < 0.001$), reflecting lower accuracy in complex compared to simple sentences. There was no significant interaction between Group and Sentence Type, which shows that both groups had lower accuracy in complex than in simple sentences.

Summary of Turkish-English findings

This study included Turkish-English sequential bilingual children and monolingual children attending schools of similar SES, as judged by the criterion of above-average percentage of free school meals. The results showed that the distribution of the monolingual children on the sentence recall task from CELF was in line with monolingual norms. However, the sequential bilingual children were less accurate than the monolingual children, and a striking 70% of the L2 children performed at least 2SD below the monolingual mean. The AoO in the L2 group, ranging from 2;5 to 5;0 years, and length of exposure, ranging from 1 to 6 years, were not found to affect performance. In considering these findings, it is important to note that none of the children were exposed to English before the age of 2;5, and they were attending schools with a high density of children speaking Turkish. Finally, sentence length and complexity affected both L1 and L2 children in a similar way. Both groups performed better on short and simple sentences than on long and complex sentences. The results of this study highlight the need for caution in interpreting results on the CELF sentence recall task in this population.

Discussion

The results of these studies of sentence repetition in L2 are strikingly divergent. The findings in the L2 Hebrew studies (with L1 Russian/English) are promising for use of sentence repetition in L2 assessment, with the majority of children (88% and 91%) performing in the normal range for monolingual children despite variations in exposure, and only 3 out of a combined group of 110 children falling more than 2 SDs below the monolingual mean. The overall distribution of performance is in line with the normal monolingual distribution. Based on these findings, any child scoring outside the normal monolingual range may be considered at risk of language impairment. However, this is not the case for the L1 Turkish/L2 English study, where the majority of L2 children (70%) scored more than 2 SDs below the monolingual mean, and only 2 of the 17 children obtained scores in the

normal monolingual range. Thus, if monolingual norms were applied, the majority of these children would be identified as impaired. The Russian-German group fell between these extremes, with 66% scoring within or above the normal monolingual range, but a notable 29% between 1 and 2 SDs below the mean, and 5% even lower.

Sources of difference

The three studies we have reported vary in multiple respects apart from language pairs and countries of residence of participants. This multiplicity of differences between studies precludes direct comparison of findings. However, analyses *within* studies indicate factors that may influence SR performance, and these throw some light on likely sources of divergence between studies. Consideration of these factors is important for drawing interim conclusions about the use of sentence repetition and for identifying further research needed to clarify the contribution of sentence repetition to assessment in L2 populations.

(i) Age of onset and length of exposure

The age range, AoO, and LoE of participants varied between and within studies, as summarised in Table 19.

Table 19: Mean (range) for age, age of onset, length of exposure in the four L2 groups (all given in months)

Study	Age	AoO	LoE
Russian-Hebrew (n=75)	70 (58-83)	35 (0-66)	37 (12-75)
English-Hebrew (n= 35)	69 (53-78)	35 (6-51)	34 (15-69)
Russian-German (n=61)	66 (47-86)	24 (0-46)	42 (13-82)
Turkish-English (n=17)	106 (73-111)	39 (29-60)	55 (21-83)

The Hebrew study found effects of AoO and a trend towards effects of LoE. It is perhaps unsurprising that children exposed between 0 and 2;0 years, who might be considered simultaneous bilinguals, perform like monolingual children. But results were good even for those exposed after age 3;0, with 82% performing in the normal range, including 13% who were more than 1 SD above the monolingual mean. In the Russian-German study, AoO and LoE did not prove significant factors (although the children scoring below -2 SD had the latest AoO and shortest LoE, and exploratory correlational analyses, not shown above, suggested a significant correlation of performance with LoE, r (60) = 0.279, p = 0.03). These factors also proved non-significant in the Turkish-English study. While the relatively small numbers in that study reduced power to identify effects, AoO and LoE within the sample varied widely, so the lack of effects is striking. Since no child in the Turkish-English study was exposed before 2;5 years, it is possible that this is a turning-point for age of onset, but our findings suggest other factors are at stake. Indices of AoO and LoE used in these studies provide a limited measure of children's language experience: following first exposure, the balance of language use at home and at school might vary widely. It is notable that the L2 children in the Turkish-English study were attending schools with a high density of Turkish-speaking children. Teasing out the effects of more subtle differences in language experience would require more detailed information about exposure and careful control of this factor.

Our studies highlight another potentially important factor which may be related to differences in language exposure, namely SES.

(ii) SES

The SES of participant groups varied between our studies. The Russian-German study included two SES groups, allowing within-study comparison of SES effects. This revealed significant differences by SES, as measured according to mothers' level of education (more vs less than 14 years of education). It is possible

that SES differences overlapped with differences in AoO and LoE, but since AoO and LoE were not significant in this study, we have some indication that SES was influential. SES might also be implicated in the very poor sentence repetition performance observed in the Turkish-English group. Although information about the SES levels of the individuals in this group was limited, all were attending schools in the low SES category according to the 'free school meals' index. In contrast, the high performing children in the Hebrew study were all from mid-to-high SES backgrounds.

SES is by no means a simple factor (Roy & Chiat, forthcoming), and whatever indices of SES are used, they may be compounded with cultural differences. Nonetheless, the suggestion that SES may be an important factor in sentence repetition performance of L2 children is in line with findings of SES effects in monolingual children (see introductory section of this chapter).

(iii) Test materials

The sentence repetition tests administered to the children differed in content and scoring. Sentence presentation is accompanied by pictures in the Hebrew test, but not in the German and English tests. The German test includes plausible and implausible sentences, while the Hebrew and English tests include only plausible targets. Sentences in the Hebrew and German tests are made up of early-acquired vocabulary, and though they contain complex structures, sentence complexity is limited. In contrast, the CELF sentence recall subtest administered to the Turkish-English group is designed for school-age children (up to age 21), and after the first six single-clause sentences, targets become progressively more complex. Almost all sentences deploy later-acquired and often school-related vocabulary. Scoring in the Hebrew and English tests is based on types and numbers of errors per sentence respectively, while the German test scores number of words correct.

The substantially greater linguistic demands of the CELF sentence recall task could be responsible for the particularly poor performance observed in the Turkish-

English group. However, comparison of L2 with L1 performance on the simple and complex targets within this task did not reveal an interaction between complexity and group, indicating that the L2 group had difficulties with simple as well as complex targets.

Implications for L2 assessment

The heterogeneous levels of L2 performance within our three studies demonstrate that the tests are tapping into children's linguistic skills. This is as we would expect from studies of sentence repetition in typically and atypically developing monolingual children, and from Polišenská's evidence that linguistic knowledge, particularly morphosyntactic and lexical knowledge, affects children's repetition capacity. Sentence repetition is therefore a useful tool for assessing linguistic *proficiency* in L2 children. As such, it has all the advantages identified for sentence repetition as a language assessment for L1 (see introductory sections). With a limited number of carefully controlled stimuli, sentence repetition is quick to administer and score and reveals not only level of language proficiency but strengths and weaknesses in targeted aspects of language. Sentence repetition tasks therefore provide an efficient and informative method for checking levels of language in L2 groups and individuals. Testing children's sentence repetition may expose low levels of language performance that have been overlooked or confused with low ability. In so doing, it would highlight ways in which school language needs to be tailored to the L2 population in the school, and indicate language needs that require additional support at a group and/or individual level. Taking the example of the Turkish-English group reported in this chapter, it is important for teachers to know that these children's ability to process sentence input in English is at a level that would be deemed impaired in their monolingual peers and to consider the support needed to raise their language to an appropriate level and facilitate their access to the school curriculum.

In monolingual children, low proficiency is assumed to reflect low ability unless there is reason to think children's language environment has been unusually limited In the case of L2 acquisition, we cannot assume that proficiency reflects ability since experience is known to vary. Individually and collectively, our studies provide extensive evidence that sentence repetition performance is heavily influenced by the heterogeneous experience of L2 children. In the Hebrew studies, AoO and LoE were found to have an effect, indicating the need to take these factors into account in judging performance. However, the Russian-German and Turkish-English studies revealed that sentence repetition may be more affected by differences in children's experience that go beyond age of onset and years of exposure and led to the suggestion that late onset and reduced exposure associated with L2 acquisition may be further compounded by limitations in language environment associated with low SES background. The disproportionate numbers of children in the Turkish-English and Russian-German studies performing in the range associated with impaired monolingual performance highlight the need for caution in using sentence repetition tests designed and normed for monolingual children for clinical assessment of L2 children.

Based on our findings, we conclude that monolingual norms on a sentence repetition task can be applied if assessing children's *proficiency* in L2. However, they cannot be applied for purposes of clinical diagnosis unless there is adequate evidence that children in the relevant L2 population with a similar level of exposure perform in line with these norms (as in the L2 Hebrew groups).

Future directions

Difficulties with morphosyntax, and specifically with function words and inflections, are a hallmark of SLI in English (Leonard, 1998), and are apparent in children's sentence repetition (Chiat & Roy, 2008; Seeff-Gabriel et al., 2008; Seeff-Gabriel et al., 2010). Although function word repetition was also found to be low in

the socially disadvantaged group studied by Roy and Chiat, this was not the case for the oldest age group, in which the rate of low performance was in line with norms (Roy & Chiat, forthcoming). As discussed above, the extent of poor performance in the Turkish-English group might in part be due to the vocabulary and syntax sampled in the CELF sentence recall test. We suggest that a test targeting more basic vocabulary and morphosyntax, while less informative about language proficiency needed for school purposes, might be more effective in distinguishing L2 children with language deficits from those with limited proficiency.

As a next step in evaluating the potential of sentence repetition in L2 clinical assessment, we propose an investigation of performance on a repetition test comprising simple sentences that are made up of early-acquired vocabulary and include representative exemplars of the morphosyntactic devices used to convey relations in the target language (word order, function words, inflections). Administration of such a test to L2 groups varying in SES as well as AoO and LoE will reveal the effects of these factors on attainment of core sentence structure and the point at which the test is valid for identifying deficits that require clinical intervention. Systematically controlled stimuli allow qualitative as well as quantitative scoring of morphosyntactic targets. This will reveal whether L2 groups show particular profiles of omission or commission errors (see Paradis, 2010; Armon-Lotem, 2010), and whether a small proportion of children lag behind or produce different errors from their L2 peers. Ideally, L2 children would be assessed using analogous sentence repetition tests in L1, to evaluate consistency of performance across languages and to determine whether children deemed at risk in one language are also found to be at risk in the other.

A programme of research along the lines we have proposed is currently being pursued under the auspices of COST Action IS0804 (http://www.bi-sli.org/). The multi-country team of researchers involved in this Action are drawing up a framework

for constructing a sentence repetition test that targets a comprehensive range of simple and complex sentence functions. Applied to particular languages, this framework can be 'filled in' and extended to ensure representative sampling of syntactic and morphosyntactic devices deployed by each language. Results of sentence repetition across a range of typologically varied language pairs will provide an extensive database for clarifying the effects of AoO, LoE, sociocultural factors, and linguistic characteristics of L1 and L2 on children's performance, and the potential for identifying children with language impairment in different languages. Once consolidated, this framework for sentence repetition will be made available. This chapter closes, then, with a promissory note founded on the theoretical case and empirical evidence we have presented.

Acknowledgements

The data collection for the English-Hebrew study was funded by the Israel Science Foundation (Grant 938). The data collection for the Russian-Hebrew and Russian-German studies was supported by the BMBF funded Consortium “Migration and societal Integration”, grant number 01UW0702B. The data collection for the Turkish-English study was supported by the Economic and Social Research Council grant number RES-061-23-0137. The writing of this paper was supported by COST Action IS0804 “Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment”.

References

Archibald, L.M., & Gathercole, S.E. (2006). Nonword repetition: A comparison of tests. *Journal of Speech, Language, and Hearing Research*, *49*, 970-983.

Armon-Lotem, S. (2010). Instructive bilingualism: Can bilingual children with specific language impairment rely on one language in learning a second one? *Applied Psycholinguistics, 31,* 253-60.

Bishop, D.V.M., McDonald, D., Bird, S., & Hayiou-Thomas, M.E. (2009). Children who read words accurately despite language impairment: Who are they and how do they do it? *Child Development, 80,* 593-605.

Brown, R. (1973). *A first language*. Harmondsworth: Penguin.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research, 40*, 519-525.

Casalini, C., Brizzolara, D., Chilosi, A., Cipriani, P., Marcolini, S., Pecini, C., Roncoli, S., & Burani, C. (2007). Nonword repetition in children with Specific Language Impairment: A deficit in phonological working memory or in long-term verbal knowledge? *Cortex, 43,* 769-776.

Coady, J.A. & Evans, J.L. (2008). Uses and interpretations of non-word repetition tasks in children with and without language impairments (SLI). *International Journal of Language and Communication Disorders, 43,* 1-40.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry, 42*, 741-748.

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, 187-208.

Dollaghan, C., & Campbell, T.F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*, 1136-1146.

Ellis Weismer, S., Tomblin, J.B., Zhang, X, Buckwalter, P., Chynoweth, J.G. & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 43*, 865-878.

Engel, P.M.J., Santos, F.H. & Gathercole, S.E. (2008). Are working memory measures free of socioeconomic influence? *Journal of Speech, Language, and Hearing Research, 51*, 1580-1587.

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of Verbal Learning & Verbal Behavior, 2,* 121-135.

Gathercole, S.E., and Baddeley, A.D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28*, 200-213.

Gathercole, S.E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27,* 513-543.

Gathercole, V.C.M. (2010). Interface of face to face? The profiles and contours of bilinguals and specific language impairment. *Applied Psycholinguistics, 31,* 282-293.

Girbau, D. & Schwartz, R.G. (2007). Non-word repetition in Spanish-speaking children with Specific Language Impairment (SLI). *International Journal of Language and Communication Disorders, 42*, 59-75.

Goralnik, E. (1995). *Goralnik Diagnostic Test*. Even Yehuda, Israel: Matan.

Graf Estes, K., Evans, J.L., & Else-Quest, N.M. (2007). Differences in the nonword repetition performance of children with and without Specific Language Impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 50*, 177-195.

Grimm, H. (2003). Sprachstandscreening für das Vorschulalter. Göttingen/Bern/Toronto/Seattle: Hogrefe.

Jones, G., Tamburelli, M., Watson, S.E., Gobet, F., & Pine, J.M. (in press). Lexicality and frequency in Specific Language Impairment: Accuracy and error data from two nonword repetition tests. *Journal of Speech, Language, and Hearing Research.*

Leonard, L.B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

Newcomer, P.L. & Hammill, D.D. (1997). *Test of Language Development. Primary (3rd edn) (TOLD-P:3)*. Austin, Texas: Pro-Ed.

Paradis, J. (2010). The interface between bilingual development and specific language impairment. *Applied Psycholinguistics, 31,* 227-252. Redmond, S.M. (2005). Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology. *Clinical Linguistics and Phonetics, 19,* 109-127.

Polišenská, K. (2011). *The influence of linguistic structure on memory span: Repetition tasks as a measure of language ability*. PhD thesis, City University London.

Polišenská, K., Chiat, S., & Roy, P. (under review). The contribution of linguistic knowledge to immediate verbal recall in children.

Redmond, S.M., Thompson, H.L., & Goldstein, S. (2011). Psycholinguistic profiling differentiates Specific Language Impairment from typical development and from Attention-Deficit/Hyperactivity Disorder. *Journal of Speech, Language, and Hearing Research, 54,* 99-117*.*

Riches, N.G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence repetition in adolescents with specific language impairments and autism: An investigation of complex syntax. *International Journal of Language and Communication Disorders, 45,* 47-60.

Roy, P. & 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*, 223-34.

Roy, P. & Chiat, S. (forthcoming). Teasing apart disadvantage from disorder: The case of poor language. To appear in C. Marshall (Ed.), *Current issues in developmental disorders* (part of the *Current issues in developmental psychology* series). London: Psychology Press.

Sahlén, B., Reuterskiöld-Wagner, C., Nettelbladt, U., & Radeborg, K. (1999). Non-word repetition in children with language impairment – pitfalls and possibilities. *International Journal of Language and Communication Disorders, 34*, 337-352.

Seeff -Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe

speech difficulties. *International Journal of Language & Communication Disorders, 45*, 691-702.

Seeff-Gabriel, B., Chiat, S., & Roy, P. (2008). *Early Repetition Battery*. London: Pearson Assessment.

Semel, E., Wiig, E., & Secord, W. (1994). *Clinical Evaluation of Language Fundamentals – Revised*. San Antonio, Texas: The Psychological Corporation.

Semel E, Wiig E, Secord W. Clinical Evaluation of Language Fundamentals (CELF-IIIUK), 3rd ed., UK. London: Psychological Corporation; 2000.

Smith, N. (1973). *The Acquisition of Phonology: A Case Study*. Cambridge: Cambridge University Press.

Stokes, S.F., Wong, A., Fletcher, P., & 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.

Wiig, E., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals – Preschool*. San Antonio, Texas: The Psychological Corporation.

Willis, C.S. & Gathercole, S.E. (2001). Phonological short-term memory contributions to sentence processing in young children. *Memory, 9,* 349-63.