J. Child Lang. **44** (2017), 1435–1457. © Cambridge University Press 2017
doi:10.1017/S0305000916000635

Assessing understanding of relative clauses: a comparison of multiple-choice comprehension versus sentence repetition*

PAULINE FRIZELLE

Department of Experimental Psychology, University of Oxford, Oxford, Oxon, UK,

and

Department of Speech and Hearing Sciences, Brookfield Health Sciences Complex, University College Cork, Cork, Ireland

CLODAGH O'NEILL

Department of Speech and Hearing Sciences, Brookfield Health Sciences Complex, University College Cork, Cork, Ireland

AND

DOROTHY V. M. BISHOP

Department of Experimental Psychology, University of Oxford, Oxford, Oxon, UK

(*Received* 9 *May* 2016 – *Revised* 16 *September* 2016 – *Accepted* 4 *December* 2016 – *First published online 16 January 2017*)

ABSTRACT

Although sentence repetition is considered a reliable measure of children's grammatical knowledge, few studies have directly compared children's sentence repetition performance with their understanding of grammatical structures. The current study aimed to compare children's performance on these two assessment measures, using a multiple-choice picture-matching sentence comprehension task and a sentence repetition task. Thirty-three typically developing children completed both assessments, which included relative clauses representing a range of syntactic roles. Results revealed a similar order of difficulty of constructions on both measures but little agreement between them when evaluating individual differences. Interestingly, repetition was the easier of the two measures, with children showing the ability to repeat sentences

[*] Address for correspondence: Pauline Frizelle, Department of Speech and Hearing Sciences, Brookfield Health Sciences Complex, University College Cork, Cork, Ireland. e-mail: p.frizelle@ucc.ie

they did not understand. This discrepancy is primarily attributed to the additional processing load resulting from the design of multiple-choice comprehension tasks, and highlights the fact that these assessments are invoking skills beyond those of linguistic competence.

INTRODUCTION

In the course of typical language development children produce relative clauses as early as around three years of age (Crain, McKee & Emiliani, 1990; Diessel & Tomasello, 2000; Jisa & Kern, 1998; Limber, 1976). However, research suggests that across languages their comprehension of the same structures does not emerge until two to three years later (de Villiers, Tager-Flusberg, Hakkuta & Cohen, 1979; Goodluck & Tavakolian, 1982; Håkansson & Hansson, 2000; Sheldon, 1974). The majority of the comprehension studies cited here use a toy manipulation paradigm where the child uses toys to act out a spoken sentence, and although significant variation is reported within their results, studies that compare production and comprehension directly (using a number of methodologies) (e.g. Håkansson & Hansson, 2000) have also reported superior production skills. This pattern of development for complex clauses contrasts with the usual finding that comprehension precedes production (Leonard, 1998), raising questions about how comprehension of these structures is assessed.

In everyday discourse, comprehension can often be achieved even if a heard sentence is only partially processed, by using context and prior knowledge to infer meaning. Formal tests of syntactic knowledge, however, typically are devised to reduce or even abolish use of context, forcing the listener to process the incoming sentence completely. Instruments have been devised to assess language comprehension by using a multiple-choice format that in effect forces the listener to form a semantic representation that relies on the syntactic structure to assign thematic roles to all the content words in a sentence. Clinical instruments typically use a one in four picture layout (one picture representing the target structure and the other three considered distractors) to reduce the probability of choosing the correct item by chance. In addition, this layout reduces the number of exemplars required to test each item effectively, thereby avoiding an assessment of unreasonable length.

Using this approach, it is possible to devise test items that can only be interpreted by those with a deep knowledge of the construction under test. At the same time, however, the multiple-choice format has the drawback that it introduces elements into the task that may lead to failure for reasons other than lack of linguistic competence. Consider the items shown in Figures 1a, 1b, and 1c. If a child is able to select the correct

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

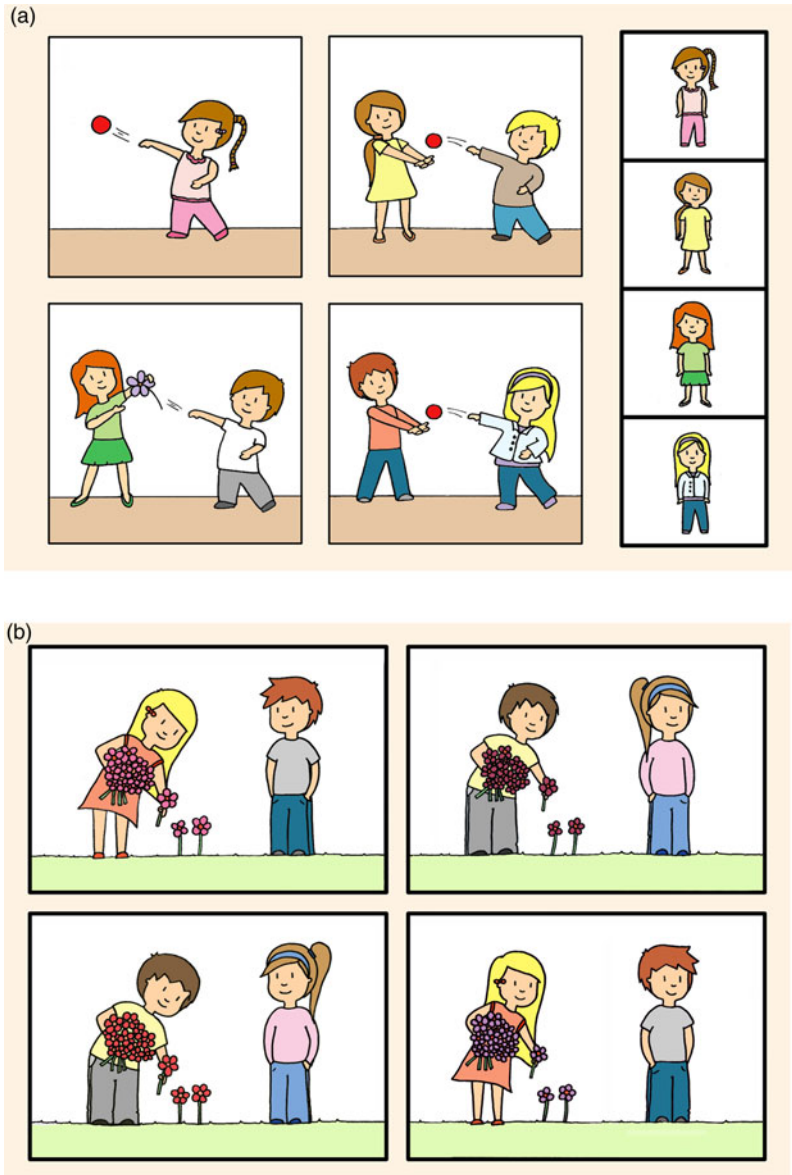

Fig. 1. (a) This is the girl he threw the ball to. (b) He saw the girl that picked the flowers. (c) The woman made the jumper that he tried on.

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

Fig. 1. (Continued)

response in a set of items such as these, this is good evidence that they are able to analyze the relative clause construction to assign thematic roles to the lexical items in the sentence. If, however, they fail, this could be due to non-linguistic factors, such as difficulty holding the sentence in memory while comparing the four pictured items, which are perceptually as well as linguistically confusing. The presence of three distractors adds a linguistic as well as cognitive load to the task. Linguistically, not only is the child required to map the semantic roles on to the syntactic structure, but they must also rule out three competing alternative mappings. The ability to rule out competing structures is likely to be influenced by other executive functions such as selective attention and inhibition.

Because of these concerns, it may be unwise to rely solely on this type of multiple-choice test to assess knowledge of complex syntax. Other approaches to assessment are possible, but each will have its own biases and complexities. For instance, we could use the method employed in most of the comprehension studies cited above, in which the task is to act out a spoken sentence. However, act-out tasks have also been criticized on the basis that they may underestimate children's knowledge due to a competing acting bias (McDaniel & McKee, 1998), i.e. children's desire to play with the toys rather than follow the instructions they hear. In addition, it has been suggested that an act-out methodology unnecessarily complicates the child's task and that many used experimentally have

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

violated appropriate pragmatic conditions by not providing a set of referents from which a subset can be distinguished (Hamburger & Crain, 1982).

One assessment method that has been widely used in recent years to assess grammatical knowledge is sentence repetition. Although at first glance this might seem to be a measure simply of the ability to repeat a string of words, a large body of research shows that this is not the case. Immediate sentence repetition has been shown to be reflective of language behaviour in natural settings (Gallimore & Tharpe, 1981), and both immediate and delayed repetition have been found to discriminate effectively between second language learners across different proficiency levels (see review by Yan, Maeda, Lv & Ginther, 2015). As far back as the late 1960s, researchers such as Slobin and Welsh (1968) and Clay (1971) argued that if sentence length exceeds an individual's short-term memory word span (the number of words they can repeat in a list), repetition will require a reliance on linguistic knowledge in long-term memory. They argued that sentence repetition reflects an individual's underlying grammatical competence, in that a person's syntactic knowledge assists them in 'chunking' components of the sentence, which facilitates the recall process. Therefore, sentences that exceed a child's short-term memory span are likely to be processed for meaning when produced successfully (Naiman, 1974; Slobin & Welsh, 1968; Vinther, 2002). More recently, Riches (2012) suggested that the roles of short- and long-term memory are not length dependent, but that they work effectively together at all sentence lengths. Researchers are now converging on the view that sentence repetition is not purely a task of reproducing a heard series of words, but that it is supported by conceptual, lexical, and syntactic representations in long-term memory (Brown & Hulme, 1995; Hulme, Maughan & Brown 1991; Klem, Melby-Lervag, Hagtvet, Lyster, Gustafsson & Hulme, 2015; Potter & Lombardi, 1990, 1998; Schweickert, 1993), as well as by phonological short-term memory processes (Alloway & Gathercole, 2005; Hanten & Martin, 2000; McCarthy & Warrington, 1987; Rummer & Engelkamp 2001).

In addition, a number of researchers have highlighted the link between sentence repetition and syntactic competence in children. Using immediate recall of subject and object relative clauses, Kidd, Brandt, Lieven, and Tomasello (2007) found that manipulating complexity while maintaining sentence length resulted in children making a greater number of sentence repetition errors. This was also the case in research carried out by Frizelle and Fletcher (2014), using the full range of relative clause types; while length remained constant, children found it increasingly difficult to immediately recall sentences as the complexity of the structure increased. The implication here is that these difficulties cannot be explained by differences in short-term phonological memory but by the underlying

syntactic competence or representations in long-term memory. The relationship between syntactic competence and sentence repetition ability is further reinforced in a recent study by Polišenská, Chiat, and Roy (2015). These authors investigated how different types of long-term linguistic knowledge contribute to children's immediate recall ability. They manipulated seven different linguistic conditions ranging from sequences of non-words to full grammatical sentences to evaluate how each condition affected children's span. They found that children's morphosyntactic knowledge played the largest role in children's immediate recall capacity: children obtained a mean span of over 3·5 words longer for grammatical than ungrammatical sentences, compared to an increased span of just 1·5 for real words vs. non-words and less than 1 for sentences that were either semantically plausible or not. Other researchers have directly compared children's performance on sentence repetition tasks to their spontaneous use of grammar. Geers and Moog (1978) reported a strong correlation between the average immediate sentence repetition error scores (from children aged four to fifteen years) and those derived from Developmental Sentence Scoring (DSS) (Lee, 1974) – a measure of grammatical complexity in spontaneous language use. In addition, McDade, Simpson, and Lamb (1982) reported a very strong correlation between the performance of four-year-old children on a sentence repetition task (with a 3-second time lapse following each sentence) and their DSS performance.

The research we have cited, coupled with the current thinking on the underlying mechanisms involved in sentence repetition, would lead us to conclude that sentence repetition is a reliable measure of children's grammatical knowledge. However, few studies have directly compared children's sentence repetition performance with their understanding of grammatical structures. The question that we consider here is how far there is agreement between sentence repetition and multiple-choice comprehension measures of competence with a type of complex sentence, relative clause constructions. There are three aspects to this question:

1. We can ask whether one task is generally easier than the other. On the one hand, we might expect a comprehension task to be easier because it does not require the child to engage language production systems, and because the presence of pictures should ease the memory load on the child, as the pictures provide a permanent concrete representation of the lexical items in the sentence. On the other hand, a repetition task does not necessitate that the sentence be assigned a semantic interpretation. Interestingly, when McDade *et al*. (1982) investigated children's sentence recall ability as a function of sentence comprehension, they found that children could repeat sentences that they did not understand. In their study, six children (aged 4;01 to 4;07) were required to repeat sixteen

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

sentences in three conditions: (a) immediate repetition followed by a request to point to the picture corresponding to that sentence (from Carrow's (1973) Test for the Auditory Comprehension of Language); (b) immediate selection of the corresponding picture followed by repetition of the sentence; and (c) repetition of the sentence following a 3-second delay. They concluded that immediate sentence recall might in fact overestimate children's language ability. However, many of the sentences could have been considered to be within children's span and the sample size was small.

2. A further question is whether the two methods agree in terms of the order of difficulty of specific constructions. If they do, then this would indicate that, despite any overall differences between repetition and comprehension, both methods are indexing a core aspect of language knowledge.
3. Children show individual variation in their task performance, raising the question of whether children vary in their syntactic competence, or whether such differences can be largely attributed to differences in non-linguistic performance factors (e.g. attention, impulsivity, etc.). We would expect performance factors to vary according to task demands, and so be different for repetition and comprehension tasks. Thus, if large individual differences are found but are consistent from task to task, this would indicate that variable syntactic competence is the main factor responsible for variation between children. If, however, there are individual differences that are inconsistent across repetition and comprehension, this would suggest that non-linguistic performance factors have a large impact on children's scores. We show here that we can distinguish these possibilities by looking at individual differences across tasks, and that it is possible to model the alternative scenarios to show which provides a better fit to observed data.

METHOD

Participants

Thirty-three typically developing children participated in the study. Of those initially recruited, two were excluded due to failing the hearing screening test, and one due to being absent on the second assessment day. The participating children were between the ages of 5;0 and 6;06 (mean age 5;07) and were recruited through primary schools in Cork city, Southern Ireland. The Cork Teaching Hospitals Ethics Committee granted written ethical approval for the study. Parents and children were required to give written consent and assent as appropriate. Children were included on the basis that they had never been referred for speech and language therapy; had typical language abilities (based on teacher and parental reports); spoke English as their first language and the language of

the home; and had no known neurological or hearing difficulties. The latter was screened for on the first day of assessment and children were required to pass three frequencies (1000 Hz, 2000 Hz, and 4000 Hz) at a 25 dB level in both ears.

Experimental tasks

Comprehension task. The comprehension task was a multiple-choice sentence–picture matching task designed to assess children's understanding of the full range of relative clause structures; subject (both intransitive and transitive), object, indirect object, oblique genitive subject, and genitive object (for a more detailed description of relative clause types see Frizelle and Fletcher, 2014). The protocol assessed fifty-six relative clauses with two matrix clause types: twenty-eight relative clauses were attached to the predicate nominal of a copular clause (containing a single proposition) and twenty-eight to the direct object of a transitive clause (full bi-clausal relatives). There were therefore four examples of each relative clause type (4×7) attached to both types of matrix clause. We included single propositional relatives, as these are the most common relative clauses to occur in young children's naturalistic speech (Diessel & Tomasello, 2000). Sentences were all between 6 and 12 syllables in length. We considered matching the stimuli on length but this would necessitate padding out some clause types with redundant words such as adjectives or adverbs. For example, it is natural that an indirect object relative would be longer than an object relative as the former contains an additional object. Given that the main aim of the study was to compare repetition and comprehension using a range of clause types, control of sentence length was not critical, so we allowed sentence length to be determined by the relative clause type. An example of the test sentences is given in Table 1.

The test sentences were chosen on the basis of previous work carried out by Diessel and Tomasello (2000, 2005) and Frizelle and Fletcher (2014) indicating a performance hierarchy in children's ability to recall these sentence structures. Based on the British National Corpus, the sentences included high-frequency nouns and verbs. These were cross-referenced with the English MacArthur Bates Communicative Development Inventory (CDI; Fenson, Marchman, Thal, Dale, Reznick & Bates, 2007) to ensure an early age of acquisition. The sentences were also modified to account for research carried out by Kidd *et al*. (2007) on the discourse regularities of young children's use of relative clauses, i.e. all object relatives had an inanimate head noun and a pronominal subject. Pronominal subjects were also used in the oblique, indirect object, and genitive relative clause structures as they were considered to be more reflective of natural discourse. Children were presented with each sentence orally and were asked to choose

TABLE 1. *Example test sentence for each relative clause type*

	Single propositional	**Bi-clausal**
Subject intransitive	This is the bird that was flying.	She followed the boy that ran.
Subject transitive	This is the girl that was drinking the milk.	He saw the girl that picked the flowers.
Object	This is the cake that they cut.	The dog ate the banana that she dropped.
Indirect object	This is the man that she poured the juice for.	He followed the girl that he gave the present to.
Oblique	This is the girl that he threw water at.	The woman made the jumper that he tried on.
Genitive subject	This is the girl whose cat caught a mouse.	He pulled the woman whose scarf was stuck.
Genitive object	This is the boy whose picture she painted.	The girl smiled at the boy whose cake she ate.

the picture (from a choice of four) that corresponded to that sentence. The other three images were distractors. The pictures were presented in three formats determined by the relative clause type. Relative clauses with a single proposition were illustrated as in Figure 1a, with a choice of four pictures, one of which represents the given sentence and the other three representing the distractors. Unlike full bi-clausal relatives, the initial verb in a single propositional relative clause usually serves as an attention getter, or in this case as a formulaic instruction to the child. In this sense one sentence is not truly embedded into the other. If given the instruction "Point to the cup that he broke", we are asking the child to point to the head noun about which the relative clause is giving more information. The more accurate response is therefore to point to 'a cup' rather than 'a man breaking a cup'. For this reason, these constructions were presented with a character or object strip of each referent (head noun) to choose from. However, if the child pointed to the main picture this was also scored as correct. Distractors for these sentences included reversed roles, verb/object distractor, or a relative clause subject distractor. Full bi-clausal relatives were represented as in Figures 1b and 1c. Structures such as that illustrated in 1c required a two-picture format, as 'the woman' needs to have made the jumper before 'he' can try it on. Distractors for the full bi-clausal relatives included role reversal of the main clause (the relative clause is understood), role reversal of the relative clause (the main clause is understood), and role reversal of both main and relative clause. These distractors are illustrated in Figures 1b and 1c.

Sentence recall task. The sentence recall task included the same sentences as those assessed in the comprehension task previously described. We decided to use live voice rather than prerecorded sentences as this helped engage these young children more readily in the task. The same examiner

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

administered both assessments with all children, ensuring a level of consistency. Children were introduced to the task as a puppet game in which they had to repeat sentences 'like a parrot'.

Procedure. Children were assessed individually in a quiet room in their respective schools. The assessments were administered in two sessions within one week of each other. The sequence of test sentences (including practice items) was randomized for both experimental tasks so that there were two orders of presentation for each task. The order in which the assessments were administered was also randomized such that half the participants completed the repetition task followed by the comprehension task, and the other half completed the tasks in the reverse order. For the multiple-choice task, children listened to a sentence and were required to point to the picture that corresponded to that sentence. For the sentence repetition task, the researcher read individual sentences and children were required to repeat them verbatim. Repetitions were allowed for both assessment protocols if background noise was evident or if it was clear that the child had not heard the test sentence properly. This resulted in a minimal number of repetitions required (less than 1% of the total number of test sentences). Positive feedback was given after each response regardless of the child's performance on either task.

The multiple-choice task was scored in real time as the children completed it. The scoring system was binary: 1 for a correct response and 0 if the response was incorrect. The sentence repetition task was recorded using a Zoom H4 audio-recorder. The responses were stored on computer for transcription and analysis. All transcriptions were orthographic. Again the scoring system was binary. Children were assigned a score of 1 if they repeated the sentence accurately or if the error made would not have resulted in an incorrect response on the multiple-choice task. For example, if the child repeated a sentence while changing definiteness, tense, or omitting an optional relativizer, this would not result in an incorrect choice on the multiple-choice task. However, if the sentence were repeated with noun or verb substitutions or omissions, noun transpositions, the omission of prepositional phrases, or as a different structure, such as coordination or another relative clause, this would result in an incorrect response in the multiple-choice task. A score of 0 was assigned in these circumstances. The rationale for using this type of scoring system was to ensure that both protocols could be compared equitably.

RESULTS

In an initial analysis, a two-way repeated measures ANOVA was used to compare the difficulty of repetition and comprehension tasks, in relation to the matrix clause type, i.e. whether sentences were single propositional (easier) or fully

bi-clausal (more difficult). The means (SDs) out of 28 for each combination were as follows: Repetition, single propositional = 26·3 (1·90); Repetition, bi-clausal = 24·9 (2·69); Comprehension, single propositional = 22·9 (2·62); Comprehension, bi-clausal = 17·6 (4·80). Because scores for the easier conditions were skewed, the standard deviations differed significantly between conditions (Levene statistic $p < .001$). To make variances more equal, total scores for all four conditions were transformed to ranks (based on all data for all conditions), after which the Levene statistic was no longer significant ($p = .140$). The transformed data were submitted to a two-way repeated measures ANOVA. This revealed a substantial effect of task type: repetition vs. comprehension ($F(1,32) = 78.1$, $p < .001$; $\eta_p^2 = .709$) and clause complexity: SP vs. bi-clausal ($F(1,32) = 48.8$, $p < .001$, $\eta_p^2 = .604$), and a significant interaction between these factors ($F(1,32) = 8.58$, $p = .006$, $\eta_p^2 = .211$).

This analysis, then, confirmed the previous finding by Frizelle and Fletcher (2014), who found single propositional constructions easier than bi-clausal constructions, and showed that this was also obtained in the comprehension task. However, in addition, and of particular interest here, was the demonstration that, first, the repetition task was much easier than the comprehension task, and second, that the difference between tasks was magnified for bi-clausal constructions.

A second question was whether the order of difficulty of constructions was similar with the two types of test. Figure 2 shows the relevant data.

A rank order was assigned to each of the 14 constructions (7 clause constructions in single propositional vs. bi-clausal form), for mean items correct in the repetition and comprehension tasks. These rank orderings were closely similar, giving a Spearman rho = .95 ($p < .001$). This result offers some evidence for the validity of the two tests as measures of language knowledge, insofar as they are both sensitive to the same aspects of clause complexity. It also suggests that the interaction between task and clause type could reflect a ceiling effect whereby the repetition task did not discriminate between clause types as many of the children found it relatively easy.

Nevertheless, if we turn to look at the extent to which there is agreement between measures in terms of estimating individual differences in children's language knowledge, the data look much less impressive. The correlation between total scores for repetition and comprehension is only .08 ($p = .656$). However, because the correlation measures only the strength of the relationship between the two variables, but not the agreement, we also completed a Bland–Altman analysis (Bland & Altman, 1995). In a Bland–Altman plot (Figure 3) the difference between the two assessment measures is plotted against the mean of the two measurements. This method allows us to calculate the mean difference between the two methods of assessment (the 'bias') and 95% limits of agreement of the mean difference (1·96 SD).

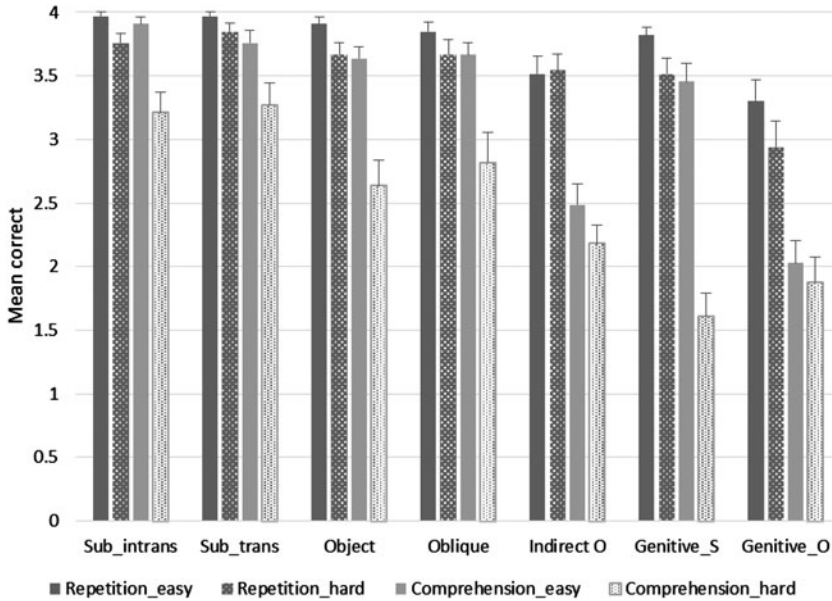

Fig. 2. Mean items correct (out of four) for repetition and comprehension, subdivided by main clause type (easy = single propositional, hard = bi-clausal) and relative clause type.

As shown in Figure 3, the two measures do not show any consistency in children's performance. There is a trend in the data, i.e. as the average performance on both measures increases, the differences between the measures are linearly decreasing, i.e. there is more agreement between the measures when children are performing at a higher level. In addition, as shown by the funnel shape of the confidence intervals, the variance around the mean difference is not constant. The results show that the lower the performance on the multiple-choice comprehension task, the more variability and the greater the differences between the two measures.

In a final analysis, we considered on an item-by-item level whether a child's knowledge as indexed by repetition agreed with their knowledge as indexed by comprehension. For each child, items were categorized as correct for both repetition and comprehension, correct for repetition and incorrect for comprehension, incorrect for repetition and correct for comprehension, or incorrect in both repetition and comprehension. For each individual, a phi coefficient and a Fisher's Exact Probability Test were computed to assess whether there was agreement on an item-by-item level. This was computed for thirty of the thirty-three children (there were 3 children who performed at ceiling on the sentence repetition task, which resulted in a zero value occupying two of the four cells of the 2×2 table,

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

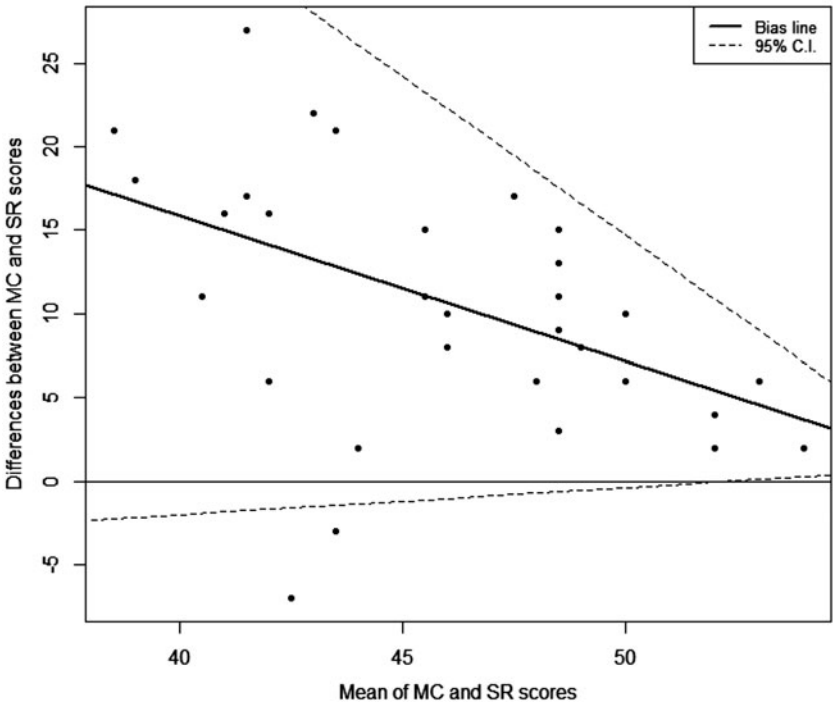

Fig. 3. Bland–Altman plot of both relative clause measures.

precluding a Fisher's exact computation). The results are shown in Table 2. With Bonferroni correction, the value required for significance was $p < .002$. There were no cases of significant agreement.

Modelling the pattern of results

Agreement between repetition and comprehension tests looks very different, depending on whether one is considering the rank ordering of difficulty of constructions, or the rank ordering of children's scores. To gain further understanding of this puzzling pattern of results, we constructed a computational model, in which different processes were simulated to see how they might affect performance.

A child's score on a given item is 1 or 0, i.e. a binary right or wrong. The factors that determine this score can be broken down into those relating to linguistic difficulty – a property of the item, which depends on its syntactic structure – and those relating to individual differences in children's syntactic competence. In addition, there will be an element of random

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

TABLE 2. *Chi-squared with continuity correction, Fisher's exact, and Phi coefficient for each child on the two assessment measures*

Child	Age	Correct in both	Correct in MC, incorrect in SR	Incorrect in MC, correct in SR	Incorrect in both	Fishers *p*	Phi
1.	6;01	37	8	5	6	0·02	0·337
2.	5;05	46	1	7	2	0·064	0·328
3	6;02	44	3	6	3	0·046	0·320
4.	5;00	32	3	14	7	0·03	0·313
5.	5;00	44	1	9	2	0·095	0·282
6.	6;02	45	0	10	1	0·196	0·273
7.	6;00	49	1	5	1	0·205	0·244
8.	5;05	39	2	12	3	0·113	0·235
9.	5;06	32	0	22	2	0·179	0·222
10.	6;06	49	2	4	1	0·249	0·204
11.	5;01	39	3	11	3	0·158	0·2
12.	6;05	34	12	5	5	0·152	0·199
13.	5;05	32	2	18	4	0·198	0·194
14.	5;01	37	1	16	2	0·239	0·176
15.	5;04	28	0	27	1	1	0·135
16.	5;05	27	3	21	5	0·451	0·132
17.	5;02	30	3	19	4	0·429	0·123
18.	5;03	37	3	14	2	0·617	0·079
19.	5;02	42	2	11	1	0·522	0·069
20.	6;01	42	2	11	1	0·522	0·069
21.	5;01	32	7	13	4	0·719	0·065
22.	5;02	30	3	20	3	0·681	0·063
23.	5;03	25	3	24	4	1	0·054
24.	6;00	35	8	10	3	0·705	0·048
25.	5;02	32	1	22	1	1	0·035
26.	5;06	31	1	23	1	1	0·028
27.	6;02	41	4	10	1	1	0·003
28.	6;01	52	1	3	0	1	−0·032
29.	5;05	41	1	14	0	1	−0·078
30.	5;04	41	2	13	0	1	−0·106

NOTES: MC – multiple-choice; SR – sentence recall; the information is given in order of the strength of the effect size; significance with Bonferroni correction: $p < \cdot002$.

error on any one trial, and on a comprehension item there will be cases where the child guesses correctly despite failing to understand.

We can simulate this situation with a formal model (see Supplementary Material, available at: <https://doi.org/10.1017/S0305000916000635>), in which the probability of success on a repetition item is:

$$R = Pi * Ps$$

Where Pi is the probability of success that reflects the linguistic difficulty of a given item, and Ps is the probability of success that reflects the syntactic competence of a specific child.

The probability of success for a comprehension item is the same, except that one in four of items that would otherwise be failed are correct by guessing, and so:

$$C = Pi * Ps + .25 * (1 - (Pi * Ps))$$

Note that the same values are used for Pi and Ps regardless of whether we are modelling comprehension or repetition. In addition, the distinction between the easy (single propositional) and difficult (bi-clausal) items for a construction is modelled by subtracting ·1 from Pi for difficult items.

Suppose we simulate the case where an item Pi is ·9 and the child's Ps is ·7. Then R is ·9 * ·7 = ·63, which means the chance of a correct response to a repetition item of this kind is ·63. C is computed as ·63 + ·25 * (1 − ·63) = ·72, so the chance of a correct response to a comprehension item of this kind is ·72.

Here we introduce a new term. 'Pc', which corresponds to variation from child to child in skills that affect comprehension only. In practice, we simulate a single child/item score by taking a value of Pc and Pi to generate values of R and C, generating a random number between 0 and 1, and assigning the item as correctly repeated if the random value falls below R, and correctly comprehended if it falls below C. We repeat this procedure for a whole set of items and children, using various ranges of Pc and Pi, to generate a simulated dataset that parallels our observed dataset. We can then compare how the simulated dataset matches the real dataset. The R script for the model is given in the Supplementary Material.

This model has seven parameters to predict: the mean scores for easy and hard repetition and comprehension items (4 parameters), the correlation between repetition and comprehension for rank ordered constructions (1 parameter), correlations between repetition and comprehension across children (1 parameter), and the average phi coefficient representing agreement between the same items for repetition and comprehension in an individual child (1 parameter). The correlations that we observe between, on the one hand, the rank ordered constructions (Figure 2), and, on the other hand, between Repetition and Comprehension scores for individual children, will depend on the range of values for Pi and Ps, and we can explore how this varies by running the simulation repeatedly with different ranges of values to see which give results that resemble those we obtained. Figure 4 shows radar charts; these are a useful way of depicting agreement between a model and obtained data when there are several different variables to consider. In Figure 4, we depict agreement between our observed results (in black) and those obtained from simulated data when different ranges of Pi and Ps are specified (in grey).[1]

[1] Correlation by structure is the correlation between the ranks of the fourteen structures for repetition and comprehension. Correlation by child is the correlation between total

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

Each spoke of this plot is a scale on which we can plot both the obtained data and predicted values for each of the seven parameters. This allows us to look at model predictions for a number of parameters simultaneously, to give a visual impression of model fit.

In Figure 4a, we show the simulated results when there is wide variation in levels of child competence (Ps range from ·6 to 1), but little variation in difficulty of the constructions (Pi range from ·95 to 1), whereas in Figure 4b these parameters are reversed. In Figure 4c, there is wide variation in both child language competence (Ps range from ·6 to 1) and difficulty of constructions (Pi range from ·6 to 1).

When there is a wide range of child ability but little variability in item difficulty (A), the simulated data do not match our results at all well, as shown by the lack of overlap between the black boundary showing obtained data and the grey area showing simulated data. This situation leads to a relatively strong correlation between repetition and comprehension across children (correlation by child), but a weaker correlation across ranked constructions (correlation by structure). The radar plot shows better agreement with the correlational data when the items have a wide range of difficulty and there is little variation in child competence (B). However, the difference in difficulty between Repetition and Comprehension items is not predicted by this model. A wide range in both item difficulty and child competence (C) again gives a poor fit – and also predicts much lower accuracy than was obtained on all item types. These simulations show that our data cannot be fit by a model that simply attributes children's success or failure on test items, to child linguistic competence and item difficulty. There must be an additional factor that can explain why comprehension and repetition do not agree well in children's individual data.

We can improve the fit of the model to the data by introducing Pc, which corresponds to variation from child to child in skills that affect comprehension only. Pc is modelled so that it exerts a greater effect on hard than easy versions of constructions, and it is uncorrelated with Ps. Figure 4d shows the situation when Pi and Ps both have a narrow high range (·95 to 1), but Pc ranges from ·6 to 1. Inclusion of this additional term improves the fit of the model to the obtained data. We retain a high correlation between rank ordering of constructions in repetition and

repetition and comprehension scores across children. Items phi is the average phi coefficient, representing correspondence of repetition and comprehension scores for individual items within children. For all these parameters, the scale is shown with a maximum score of 1 and minimum of −1. The other four parameters are the mean overall scores for Easy and Hard items on Repetition and Comprehension. Here the scale ranges from 14 to 28. A good-fitting model should give a reasonable match for all these parameters, and so overlap with the area defined by the bold line, which corresponds to obtained values.

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

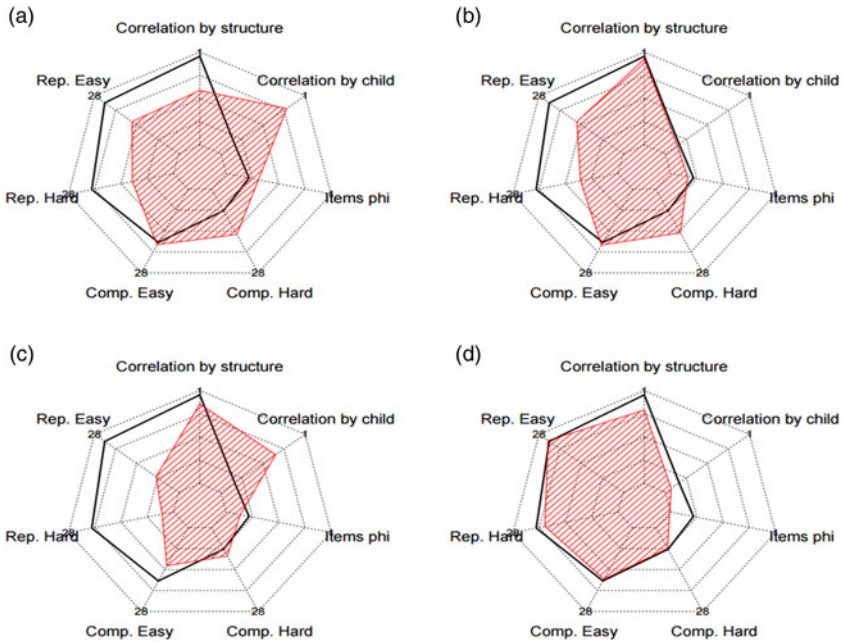

Fig. 4. Radar charts showing obtained values for different parameters of the model.

comprehension, and a low correlation between repetition and comprehension scores for individuals, while achieving a better estimation of the pattern of mean scores for easy and hard items in repetition and comprehension.

The model thus clarifies formally an intuitive explanation for the pattern of data, which is that children's performance on the comprehension task is affected by factors other than syntactic competence, which show fairly wide variation between children and can therefore lead to a lack of consistency between repetition and comprehension.

It is also worth noting that in no case does a simulation give a high value for the average phi coefficient, which reflects concordance between repetition and comprehension at the item level. For individual items, chance plays a role in determining scores, and it is clear that, even when there are strong effects of grammatical structure on item difficulty, we cannot expect a high agreement between individual items across repetition and comprehension.

DISCUSSION

The current study aimed to investigate the level of agreement between two assessment measures commonly used clinically and in child language research (a multiple-choice picture-matching sentence comprehension task

https://doi.org/10.1017/S0305000916000635 Published online by Cambridge University Press

and a sentence repetition task). In addition, we aimed to explore whether one task was generally easier than the other and whether the two methods would agree in terms of the order of difficulty of specific constructions. The results revealed that the repetition task was the easier of the two assessment tasks, in that many of the children showed the ability to repeat sentences that they did not understand when tested on the multiple-choice comprehension task. This is particularly thought-provoking in the context of current thinking regarding the need to process a sentence for meaning before reproducing it in recall. In addition, both tasks revealed a similar order of difficulty of constructions, providing some validity for what the two methods are measuring. However, despite this, when we looked at the two tasks in relation to measuring individual differences, there was very little agreement between them.

In interpreting our results it may be prudent to appraise what exactly we are assessing when administering the type of comprehension task described. Our intuition was that the comprehension test involved skills over and above language knowledge; our modelling of test performance was conducted to clarify whether the overall pattern of results would be compatible with such an interpretation, and confirmed that it was. This led us to consider what is driving performance on this type of assessment, other than grammatical knowledge. In this task the test design is such that the sentence distractors are increasing the processing load considerably when attempting to understand each sentence presented. If we consider the test sentence *He saw the girl that picked the flowers* (shown in Figure 1b), the distractor images reflect the sentences *The girl that picked the flowers saw the boy, The boy that picked the flowers saw the girl*, and *She saw the boy that picked the flowers*. Each distractor is providing an alternative regarding 'who did what to whom', and is requiring the child to process each component of the sentence in a way that would not be necessary if the same sentence were used in natural discourse. Gennari and MacDonald (2008) found that, in a group of adults, relative clause comprehension difficulty was connected to their beliefs about how the structure and thematic roles would be assigned in a given sentence. In the current comprehension task the child is required to listen to the sentence, using world knowledge and the distributional regularities of the input. We assume that they then make some kind of prediction based on typical thematic role to verb argument mapping. However, it seems that the alternative mappings that are presented in the distractors are significantly increasing the processing load in relation to what the child is trying to comprehend. Difficulty resolving structural ambiguity is often attributed to the competition of alternative interpretations, and although relative clauses are considered to be structurally unambiguous (Gibson, 1998), it could be argued that, by presenting the distractors in the manner outlined,

we are creating a level of ambiguity in thematic role assignment. As the child attempts to process the sentence, the distractors serve to activate other structures within the relative clause family, and the child is required to rule out three competing alternative interpretations.

In addition, Gennari and MacDonald (2009) noted that, when thematic roles assigned by the verb could be applied to either noun, participants had greater comprehension difficulty. Although this was not the case for the object relatives in our comprehension task (our head noun was always inanimate, a reflection of discourse regularities), in the majority of the target full bi-clausal sentences and their corresponding distractors, the thematic role assigned could be applied to either noun. This is also evident in many other assessments where complex sentences are being receptively assessed (e.g. Test for the reception of grammar (TROG; Bishop, 2003); Clinical Evaluation of Language Fundamentals (CELF – 4; Semel, Wiig & Secord, 2006). It is therefore likely that the design of the distractors is overloading the language processor and inflating the receptive difficulty level of each sentence. Moreover, the requirement to remember the given structure while being faced with three competitors, using similar nouns to the target sentence, but where the structure and thematic roles are assigned differently, is creating a significant working memory load. The distractors serve to compete and interfere with each other in memory, causing the task to be influenced to a greater degree by other cognitive functions. This influence of similarity-based interference on sentence comprehension has been noted by a number of working memory researchers (e.g. Gordon, Hendrick & Johnson, 2001, 2004; Van Dyke, 2007), particularly when there is syntactic or semantic overlap with the distractors available in working memory. The influence of the distractors on task performance is highlighted in the fact that the impact of the matrix clause type was especially marked in the multiple-choice comprehension task. Specifically, there was a bigger discrepancy between children's performance on the sentence recall vs. the multiple-choice task when the sentences were fully bi-clausal than when they had only a single proposition. The nature of single propositional vs. full bi-clausal relatives is such that the former require a different distractor set (not all relating to thematic role assignment), whereby the processing load is reduced. If we consider the single propositional sentence *This is the woman that kissed the baby*, the distractors for this sentence type include a verb (*the woman that* ***held*** *the baby*) an object (*the woman that kissed the* ***man***), and a role reversal (*the* ***baby*** *that kissed the* ***woman***). In addition, those sentences in which the head noun is inanimate do not include any distractors relating to thematic role assignment, reducing the processing load even further. The influence of task design is also evident in the subject genitive comprehension results. As shown in Figure 2, there is a particularly large discrepancy between children's comprehension of the single propositional vs. the full bi-clausal

constructions on this relative clause type. Having analyzed the design of the pictures, it became apparent that, in relation to the single propositional subject genitive relatives, children could in fact choose the correct picture without fully understanding the genitive aspect of the relative clause. In order to fully assess children's understanding of this construction, an additional referent is required in each distractor, hence how the distractors were depicted influenced children's performance on the task. Indeed, when using this type of assessment design, it seems we are increasing the influence of non-syntactic factors, while assessing some kind of absolute understanding in a context devoid of ecological validity.

An alternative interpretation of our results focuses on sentence repetition as a measure of language knowledge – it may be that a sentence repetition task is simply a better indicator of children's grammatical knowledge than a multiple-choice comprehension task. However, if children are showing the ability to repeat sentences that they cannot understand, using some kind of rote repetition, this leads us to question what is supporting their recall beyond phonological short-term memory. Previous discussions regarding children's ability to reproduce sentences by rote repetition have centred on whether (i) the length of the input is within span or (ii) the complexity of the sentence is beyond the child's grammatical knowledge. Either of these scenarios will result in children relying heavily on their phonological short-term memory in order to repeat a sentence. In the first scenario, it is argued that, if within span, the child could repeat the sentence as they would a string of unrelated words (primarily using their phonological short-term memory). In the second scenario, without sufficient grammatical knowledge the child is forced to rely on the acoustic information without decoding the sentence structure for meaning (again relying heavily on phonological short-term memory). In both these scenarios, to aid recall the child can tap into their vocabulary knowledge (stored in long-term memory), which will be influenced by a number of factors, such as word frequency, neighbourhood density, and imageability. However, the phonological short-term memory load remains high. In contrast, recalling a sentence with comprehension involves the child accessing their syntactic knowledge in long-term memory to facilitate their understanding – the meaning of the sentence and the child's linguistic knowledge are central to their ability to reconstruct the sentence for repetition. When recalling a sentence with comprehension, the role of phonological short-term memory is therefore somewhat diminished. Our results suggest that perhaps the distinction often made between the ability to repeat by rote and to repeat with full comprehension is not a helpful one. Indeed, Yan *et al*. (2015) posit that these two skills could be regarded as two extremes on a continuum. If we apply a usage-based model to language learning, the emphasis is on input frequency and distributional

learning (Tomasello, 1992), both of which impact young children's mental representations and linguistic knowledge. Using a frame and slot account, the statistical information derived from the syntactic frames and the lexical slots within them, allows children to predict words or syntactic chunks as they process a given sentence. Potentially this would allow children to repeat at least part of a sentence by rote without fully understanding it. This model has been put forward in relation to morphological learning (Lieven, Pine & Baldwin, 1997). Using a distributional approach, it is suggested that in early language development children begin to use morphemes without understanding their meaning or their grammatical relationship with related morphemes. An application of this model to syntax would account for some recall ability without complete understanding, which would go some way towards explaining the results of the current study.

To conclude, when using either a sentence repetition or a multiple-choice sentence comprehension task, it may be more helpful to consider language knowledge as involving a spectrum of abilities, rather than an absolute level of understanding. It would be of interest to examine how varying the context would alter children's receptive performance on a given sentence, with sentence repetition as one contextual representation and a multiple-choice sentence–picture matching task as another. Moreover, it may be useful for clinicians to assess children's syntactic knowledge using both methods of assessment, while being aware that the typically administered multiple-choice comprehension task is invoking other skills beyond those of linguistic competence. This is particularly relevant for researchers who are exploring relationships between measures of language comprehension and executive functioning, and may have implications for the validity of their results.

SUPPLEMENTARY MATERIALS

For supplementary material for this paper, please visit: <https://doi.org/10.1017/S0305000916000635>.

REFERENCES

Alloway, T. P. & Gathercole, S. E. (2005). Working memory and short-term sentence recall in young children. *European Journal of Cognitive Psychology* **17**, 207–20.

Bishop, D.V.M. (2003). The Test for Reception of Grammar. *TROG 2*. London: Psychological Corporation.

Bland, J. M. & Altman, D. G. (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* **346**, 1085–7.

Brown, G. D. A. & Hulme, C. (1995). Modeling item length effects in memory span: No rehearsal needed? *Journal of Memory and Language* **34**, 594–621.

Carrow, E. (1973). *Test for the Auditory Comprehension of Language*. Austin, TX: Learning Concepts.

Clay, M. M. (1971). Sentence repetition: elicited imitation of a controlled set of syntactic structures by four language groups. *Monographs of the Society for Research in Child Development* **36**, 1–85.
Crain, S., McKee, C. & Emiliani, M. (1990). Visiting relatives in Italy. In L. Frazier & J. de Villiers (Eds.), Language processing and language acquisition. Dordrecht: Kluwer.
De Villiers, J., Tager-Flusberg, H. B., Hakuta, K. & Cohen, M. (1979). Children's comprehension of relative clauses. *Journal of Psycholinguistic Research* **8**, 499–518.
Diessel, H. & Tomasello, M. (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics* **11**(1/2), 131–51.
Diessel, H. & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language* **81**(4), 1–25.
Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S. & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: user's guide and technical manual*. Baltimore, MD: Paul H. Brookes.
Frizelle, P. & Fletcher, P. (2014). Relative clause constructions in children with specific language impairment. *International Journal of Language & Communication Disorders* **49**, 255–64.
Gallimore, R. & Tharp, R. G. (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning* **31**(2), 369–92.
Geers, A. E. & Moog, J. S. (1978). Syntactic maturity of spontaneous speech and elicited imitations of hearing impaired children. *Journal of Speech and Hearing Disorders* **43**, 380–91.
Gennari, S. P. & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language* **58**, 161–87.
Gennari, S. P. & MacDonald, M. C. (2009). Linking production and comprehension processes: the case of relative clauses. *Cognition* **111**, 1–23.
Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* **68**(1), 1–76.
Goodluck, H. & Tavakolian, S. (1982). Competence and processing in children's grammar of relative clauses. *Cognition* **11**, 1–27.
Gordon, P. C., Hendrick, R. & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition* **27**, 1411–23.
Gordon, P. C., Hendrick, R. & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language* **51**, 97–114.
Håkansson, G. & Hansson, K. (2000). Comprehension and production of relative clauses: a comparison between Swedish impaired and unimpaired children. *Journal of Child Language* **27**, 313–33.
Hamburger, H. & Crain, S. (1982). Relative acquisition. In S. Kuczaj (Ed.) *Language development, vol. 1: syntax and semantics* (pp. 245–274). Hillsdale, NJ: Erlbaum.
Hanten, G. & Martin, R. C. (2000). Contributions of phonological and semantic short-term memory to sentence processing: evidence from two cases of closed head injury in children. *Journal of Memory and Language* **43**(2), 335–61.
Hulme, C., Maughan, S. & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language* **30**(6), 685–701.
Jisa, H. & Kern, S. (1998). Relative clauses in French children's narrative texts. *Journal of Child Language* **25**, 623–52.
Kidd, E., Brandt, S., Lieven, E. & 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**(6), 860–97.
Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S.-A. H., Gustafsson, J. E. & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science* **18**(1), 146–54.

Lee, L. (Ed.) (1974). *Developmental sentence analysis: a grammatical assessment procedure for speech and language clinicians*. Evanston, IL: Northwestern University Press.
Leonard, B. L. (1998). The language characteristics of SLI: a detailed look at English. In B. L. Leonard (ed.), *Children with Specific Language Impairment* (pp. 53–94). Cambridge, MA: MIT Press.
Lieven, E. V. M., Pine, J. M. & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language* **24**, 187–219.
Limber, J. (1976). Unraveling competence, performance and pragmatics in the speech of young children. *Journal of Child Language* **3**, 309–18.
McCarthy, R. & Warrington, E. (1987). The double dissociation of short-term memory for lists and sentences. *Brain* **110**(6), 1545–63.
McDade, H. L., Simpson, M. A. & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar: a question of validity. *Journal of Speech and Hearing Disorders* **47**(1), 19–24.
McDaniel, D. & McKee, C. (1998). *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.
Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism* **2**, 1–37.
Polišenská, K., Chiat, S. & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language & Communication Disorders* **50**(1), 106–18.
Potter, M. C. & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language* **29**, 633–54.
Potter, M. C. & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language* **38**, 265–82.
Riches, N. G. (2012). Sentence repetition in children with specific language impairment: an investigation of underlying mechanisms. *International Journal of Language & Communication Disorders* **47**(5), 499–510.
Rummer, R. & Engelkamp, J. (2001). Phonological information contributes to short-term recall of auditorily presented sentences. *Journal of Memory and Language* **45**(3), 451–67.
Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory and Cognition* **21**, 168–75.
Semel, E., Wiig, E. M. & Secord, W. (2006). *Clinical Evaluation of Language Fundamentals—Fourth Edition, UK Standardisation (CELF–4 UK)*. London: Pearson Assessment.
Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior* **13**, 272–81.
Slobin, D. I. & Welsh, C. A. (1968). Elicited Imitation as a research tool in developmental psycholinguistics. *Working Papers of the Language Behavior Research Laboratory*, University of California, Berkeley, 10.
Tomasello, M. (1992). *First verbs: a case study of early grammatical development*. Cambridge: Cambridge University Press.
Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory and Cognition* **33**, 330–407.
Vinther, T. (2002). Elicited imitation: a brief overview. *International Journal of Applied Linguistics* **12**(1), 54–73.
Yan, X., Maeda, Y., Lv, J. & Ginther, A. (2015). Elicited imitation as a measure of second language proficiency: a narrative review and meta-analysis. *Language Testing* **33** (4), 497–528.