LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Assessing Children's Understanding of Complex Syntax: A Comparison of Two Methods

Pauline Frizelle [a,b], Paul Thompson,[a] Mihaela Duta,[a] and Dorothy V. M. Bishop[a]

[a]University of Oxford and [b]University College Cork

We examined the effect of two methods of assessment—multiple-choice sentence–picture matching and an animated sentence-verification task—on typically developing children's understanding of relative clauses. A sample of children between the ages of 3 years 6 months and 4 years 11 months took part in the study ($N = 103$). Results indicated that (a) participants performed better on the sentence-verification than on the multiple-choice task independently of age, (b) each testing method revealed a different hierarchy of constructions, and (c) the impact of testing method on participants' performance was greater for some constructions than others. Our results suggest that young children can understand complex sentences when they are presented in a manner that better reflects how people process language in natural discourse. These results have

This research was supported by funding to the first author from the charity RESPECT and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013), under REA grant agreement no. PCOFUND-GA-2013-608728. The fourth author was supported by a Principal Research Fellowship (082498/Z/07/Z) from the Wellcome Trust. We are very grateful to the parents, children, schools, and nurseries that participated in this study. Thanks are also due to Edward Hogan and Johanne Nedergaard for assistance with data collection. For Kuppu, my friend and colleague, RIP.

This article has been awarded an Open Data badge. All data are publicly accessible via the Open Science Framework at https://osf.io/27gwb. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Correspondence concerning this article should be addressed to Pauline Frizelle, Department of Speech and Hearing Sciences, Brookfield Health Sciences Building, University College Cork, College Road, Cork, Ireland. E-mail: p.frizelle@ucc.ie

DOI: 10.1111/lang.12332

implications for the study of language comprehension in suggesting that results from multiple-choice tasks may not generalize to other methods.

Keywords complex syntax; relative clause; language; development; children; assessment

Introduction

During the process of child language development, it is often the case that children will demonstrate language knowledge that is sufficient to support comprehension but that is insufficient for production (Fraser, Bellugi, & Brown, 1963). This is not surprising, given the extra demands of utterance planning and sentence formulation involved in language production. However, in the case of complex syntax, the opposite appears to be the case. Although young children use complex sentences at a very early age (Diessel, 2004), they do poorly on tests designed to assess their understanding of them (Clark, 1971; de Villiers, Tager-Flusberg, Hakuta, & Cohen, 1979; Goodluck & Tavakolian, 1982; Hamburger & Crain, 1982; Kidd & Bavin, 2002; Sheldon, 1974; Tavakolian, 1977). Many of these studies were conducted several decades ago, and they have been criticized for the methods used (e.g., McDaniel & McKee, 1998). In addition, the constructions being tested were considered very different from those that young children say and hear in natural discourse (Diessel & Tomasello, 2005); consequently, these previous findings may bear little relation to language comprehension in naturalistic settings. The results of the majority of these studies may therefore not reflect what children are capable of understanding in relation to complex sentences, raising questions about how best to assess receptive syntactic knowledge.

There are a number of paradigms typically used to assess children's comprehension of syntax, most of which share commonalities in relation to assessment principles. Commonalities include the setup, which is usually purposely structured to remove context or situational cues in order to evaluate understanding of the specific aspect of language of interest, such as the grammatical structure, morphological marker, or a combination thereof. In a typical test (administered clinically), the items (e.g., sentences) are sequenced so that they make increasing demands on processing skills in order to identify the upper threshold of a child's comprehension ability. This is evident in the Test for Reception of Grammar (Bishop, 2003), the Clinical Evaluation of Language Fundamentals–Fourth Edition UK (CELF-4; Semel, Wiig, & Secord, 2006), the Token Test for Children–Second Edition (McGhee, Ehrler, & DiSimoni, 2007), and the Assessment of Comprehension and Expression 6–11 (Adams,

Coke, Crutchley, Hesketh, & Reeves, 2001). Despite these commonalities, there are also many differences in the assessment methods and procedures used to investigate children's comprehension of syntax, such as preferential looking, act-out tasks, story comprehension, grammaticality judgments, online methods, picture-selection tasks, and truth–value judgment tasks. Of interest in this study is the effect that different methodologies can have on the language knowledge being measured. Is it that children do not have the knowledge, or is it that they are unable to display it due to the demands of the task? Given the variation in assessment procedures used, it is likely that they will have a varied effect on the knowledge being tested and consequently on the outcome measure.

Assessment of Children's Syntax

Although each assessment measure comes with its own biases, one of the main characteristics of a good test is that the methodology should not confound the knowledge being measured with other factors. It would be misleading for both the research and clinical communities to use assessments that yield results reflecting the testing method rather than knowledge of the item being measured. In a recent study, Frizelle, O'Neill, and Bishop (2017) compared children's performance on a sentence recall task and a multiple-choice comprehension task of relative clauses. The authors found that, although the two assessments revealed a similar order of difficulty of constructions, there was little agreement between the two for evaluating individual differences. In addition, children showed the ability to repeat sentences that they did not understand. Given that sentence repetition has been reported to be a reliable measure of children's syntactic knowledge (Gallimore & Tharpe, 1981; Kidd, Brandt, Lieven, & Tomasello, 2007; Polišenská, Chiat, & Roy, 2015), the authors attributed the discrepancy between the two measures to the additional processing load resulting from the design of multiple-choice picture-selection comprehension tasks.

Multiple-Choice Picture-Selection Task

Here we consider two methods of receptive language assessment—the multiple-choice picture-selection task and a newly devised truth–value-judgment/sentence-verification animation task. The former is one of the most commonly used assessment paradigms for evaluating children's understanding of syntax within a formal, standardized assessment framework; it is used, for example, in the CELF-4 (Semel et al., 2006), Test for Reception of Grammar (Bishop, 2003), and the Assessment of Comprehension and Expression 6–11 (Adams et al., 2001). In a picture-selection paradigm, a child is presented with a sentence and asked to select the picture (usually from a choice of four) that

best corresponds to the stimulus presented. Within a four-picture layout, one picture represents the target structure and the other three are considered distractors. On the condition that each distractor is equally plausible, this reduces the probability of choosing the correct item by chance to 25%. However, it may not always be possible to design three equally plausible distractors (Gerken & Shady, 1996), in which case deciding what constitutes chance performance becomes difficult. Not only is the semantic plausibility of the distractors influential but also their syntactic framework and the relationship between the two can influence performance. In parsing a sentence for syntactic comprehension, a child will typically form a semantic representation of that sentence, relying on the syntactic structure to assign the thematic roles to the appropriate words in the sentence. However, there are additional linguistic factors that influence a child's performance, and these are not always controlled for. For example, greater comprehension difficulty has been noted when the thematic roles assigned by the verb could be applied to either noun (Gennari & MacDonald, 2009).

It is also the case that on hearing the target stimulus, children will use typical thematic role-to-verb argument mapping in trying to process the given structure. The presence of three distractors requires children to rule out three competing alternative mappings, increasing the processing load considerably and creating a level of ambiguity in how the roles should be assigned. For example, in the test sentence "She pushed the man that was reading the book" used in the multiple-choice task and shown in Figure 1, the distractor images represent the sentences "The woman that was reading the book pushed the man," "The man pushed the woman that was reading the book," and "The man that was reading the book pushed the woman." Using the same lexical set, the four pictures represent four alternative ways to assign the thematic roles. With regard to relative clauses (see Table 1), this type of distractor design is used in commercially available standardized assessments such as the Test for Reception of Grammar (Bishop, 2003) and the Assessment of Comprehension and Expression 6–11 (Adams et al., 2001). Interestingly, in the example taken from the CELF-4, the object (the banana) is inanimate and therefore the roles of the subject and object cannot be reversed, necessitating a different set of distractors. However, the distractors in the example taken from the CELF-4 shown in Table 1 serve to test children's understanding of prepositions and, on condition that the children understand the concept of *under*, they could choose the correct item. In addition, in the illustrations depicting the distractors, there is no alternative to the head noun to which the relative clause was referring (no other boy), making the use of a relative clause pragmatically inappropriate.

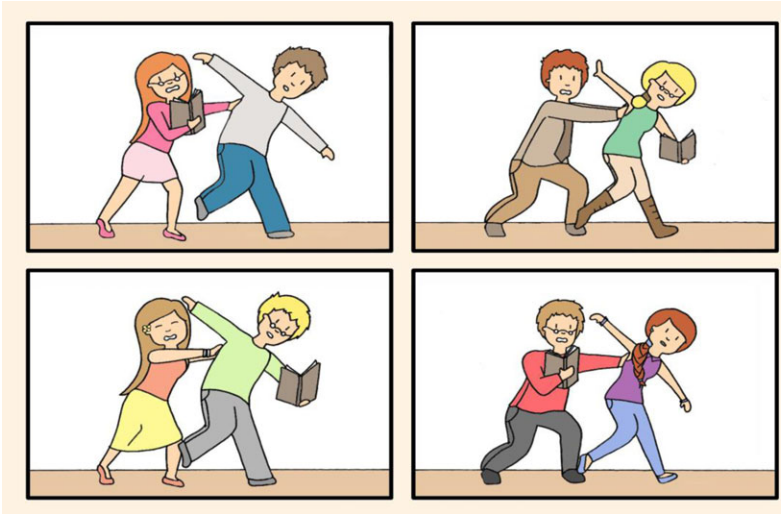

Figure 1 Example of an illustration for the test sentence "She pushed the man that was reading the book." [Color figure can be viewed at wileyonlinelibrary.com]

Table 1 Examples of distractors used in commercially available standardized assessment tests

Assessment of Comprehension and Expression 6–11	
Target Sentence	The cat that scratched the fox is fat.
Distractors	The cat that the fox scratched is fat.
	The fox that scratched the cat is fat.
	The fox that the cat scratched is fat.
Test for Reception of Grammar–Version 2	
Target Sentence	The girl chases the dog that is jumping.
Distractors	The dog chases the girl that is jumping.
	The girl that is jumping chases the dog.
	The dog that is jumping chases the girl.
Clinical Evaluation of Language Fundamentals–Fourth Edition UK	
Target Sentence	The boy who is sitting under the big tree is eating a banana.
Distractors	The boy who is standing beside the small tree is eating a banana.
	The boy who is sitting beside a small tree is eating a banana.
	The boy who is sitting in a big tree is eating a banana.

The advantage of using this role-reversal distractor approach is that researchers can devise test items that can only be correctly identified by children with a deep understanding of the syntactic structure being tested. One could also argue that, through careful manipulation of the distractor images, researchers can complete a systematic error analysis, which will potentially shed light on children's developing representations. However, at the same time, this method necessarily assesses understanding in a way that is far removed from natural discourse. There is a risk that error patterns may reflect the plausibility of particular interpretations or the salience of distractors rather than underlying linguistic representations. In addition, this format places demands on attention and memory and is likely to lead to children's failing for reasons other than a lack of linguistic knowledge (Frizelle et al., 2017). Thus, the multiple-choice picture-selection task is an example of an assessment procedure in which the method is likely to have a strong effect on the knowledge being measured.

Truth–Value-Judgment/Sentence-Verification Task

An alternative assessment measure is the truth–value-judgment/sentence-verification task, described by Gordon (1996, p. 211) as "one of the most illuminating methods of assessing children's linguistic competence." A truth–value-judgment task requires children to make a binary judgment about whether a statement accurately describes a situation presented in a given context. Some studies have used a yes/no version within the context of a short story, that is, children are presented with a story followed by several yes/no questions (see Gordon & Chafetz, 1986). Other researchers such as Crain and McKee (1985) have used what is termed a reward/punishment version. In this design, Crain and McKee asked children to watch a situation acted out while it was simultaneously described. Following a statement made by Kermit the Frog about an event, children were asked to reward him if it was true and to punish him by feeding him a rag if it were false.

In this study, we used a sentence-verification task that could be considered somewhat less demanding than those previously described. In this task, participants listen to a single sentence while simultaneously being shown an animation and must simply decide if the sentence is truly represented in the animation shown. Half of the sentences are accurately represented by the animation and half are not. By listening to the stimulus sentence at the same time as it is acted out in real time, participants can take advantage of a reduction in memory demands. However, one disadvantage of using this method is that a larger number of stimuli are required to determine if performance is reliably

above chance, which may result in fatigue, particularly for young children. In addition, Crain, Thornton, Boster, and Conway (1996) raised some methodological design concerns in relation to single sentence truth–value judgment tasks and the pragmatic context in which they are represented. These concerns were prompted by studies by Philip (1991, 1992) and Roeper and de Villiers (1993) showing that, when children were interpreting sentences containing universal qualifiers (such as every), they were influenced by the extra elements in the picture. Children showed a pragmatic bias, surmising that the extra element was there for a reason and therefore responded negatively and incorrectly to the question asked. Crain et al. followed up with further research in which they created a story context where children could plausibly reject the extra elements depicted. They argued that, with the appropriate contextual support, children would not override their grammatical knowledge in favor of any pragmatic bias. However, they also acknowledged that children's response to the number of items in each illustration could have as much to do with the foregrounding or backgrounding of the items as with the pragmatic context. For instance, visual salience depends on a number of factors such as color, size, and the relative positioning of the people/objects in the illustration. Although researchers have previously discussed this in relation to universal qualifiers, it is relevant to any truth–value judgment task where children are presented with a single image or animation.

Even when an image is an accurate representation of the sentence, the extra elements within that image are in effect a form of distractor. On the other hand, if an image does not represent the sentence accurately, the entire image is considered a distractor, and how that is structured affects the processing load for children. For example, in a distractor animation for the relative clause "She pushed the boy that was jumping," children would be shown a sequence where there are two potential referents—a boy that is jumping and a boy that is not—and the animation would show a girl pushing the boy that is not jumping. If, however, the girl is also jumping, this would increase the salience of an erroneous interpretation of the sentence "The girl that was jumping pushed the boy," which in turn increases the difficulty level of the item. Assessors aim to gain the most accurate picture of children's receptive syntactic knowledge while at the same time gaining some insight into the difficulties that they experience. However, we do not want to distract or provoke children into making an invalid response. Therefore, how each structure is represented is central to the assessment results that are likely to be obtained.

The advantage of using a truth–value-judgment task where children are shown one animation at a time is that they can evaluate the truth of each sentence

(as it is happening) directly against the real-world situation without having to store in memory the arguments associated with the verbs. In this regard, the task has no greater memory load than that required for processing everyday discourse. Children are required to make a semantic evaluation of a sentence and to map the thematic roles to the appropriate verb argument structures without having to actively rule out three competing alternative mappings. In addition, using animation reduces the dependence on children's ability to interpret adult-created pictures and reduces the amount of inference making required from them. One might therefore expect this assessment framework to yield more valid test results than those obtained from the multiple-choice picture-selection task.

The Current Study

In this study, we used the two methods of assessment previously described to measure the same language knowledge (i.e., child participants' understanding of relative clauses) to examine the extent to which assessment method affects assessment outcome. We then investigated if the use of these two different methods significantly affected the participants' understanding of restrictive relative clauses. The types of relative clauses studied here were the first four in the noun phrase accessibility hierarchy (Keenan & Comrie, 1977): subject (transitive ST and intransitive SI), object (Obj), indirect object (IO), and oblique (Obl). The labels relate to the item in the sentence that is relativized. Previous developmental work had indicated that this sequence is reflected in acquisitional order in spontaneous speech data (Diessel, 2004). In addition, we explored the relationship between children's performance on the two measures of syntactic comprehension (multiple-choice picture-selection and sentence-verification tasks) and on an independent measure of language—a sentence-recall task—used as a validation measure at it is considered to reflect an individual's underlying syntactic competence (Clay, 1971; Slobin & Welsh, 1968). The following research questions were addressed:

1. Do children show greater understanding of relative clauses when assessed using a sentence-verification animation task or a multiple-choice picture-selection task?

Based on findings reported in Frizelle et al. (2017) showing that the multiple-choice picture-selection task assesses factors other than those that are linguistic, we predicted that participants would achieve higher scores on the sentence-verification task than on the multiple-choice task.

2. Is the hierarchy of structures similar across the tests, and does it reflect the noun phrase accessibility hierarchy?

Given that the exact same sentences were used in both assessments (with additional sentences added to the sentence-verification task), we might predict that a similar hierarchy would be shown by both assessment measures.

3. What is the relationship between children's performance on the two methods of syntactic assessment and their language skills as measured by a sentence-recall task?

If our sentence-verification task is a less compromised measure of syntactic comprehension and sentence recall is a reliable measure of syntactic knowledge (Frizelle & Fletcher, 2014a; Gallimore & Tharpe, 1981; Geers & Moog, 1978; Kidd et al., 2007; McDade, Simpson, & Lamb, 1982; Polišenská et al., 2015), then we would expect the sentence-verification task to be more closely associated with the sentence-recall measure (used here as an additional measure of language knowledge).

Method

Participants

A total of 110 typically developing children between the ages of 3 years 6 months and 5 years 0 months were recruited for the study. We chose this age range based on (a) pilot work where we found that the majority of children between the ages of 3 years and 3 years 5 months engaged in guessing and (b) the age at which one expects reasonable expressive knowledge of these structures to be present (Diessel & Tomasello, 2000; Jisa & Kern, 1998; Limber, 1976). Seven children were subsequently excluded as a result of speech and language delay (3), attention difficulties (2), failure on the hearing screening tests (1), and an unwillingness to participate (1). Those with speech and language delay were diagnosed on the basis that they were attending their local speech and language therapy services. This resulted in 103 children participating in the study. The sample size was calculated using a power analysis conducted by simulation ($N = 100$). The script for the calculation and all other statistical analyses are available via Open Science Framework at https://osf.io/27gwb. The simulate function from the R package glmer (Bates, Maechler, Bolker, & Walker, 2015) was used to generate data using the parameters from the linear mixed-effects model. Further details of the approach used can be found at

https://rpubs.com/bbolker/11703. To take account of the rapid changes in language development in this age range, we divided the participants into six age bands, with between 16 and 18 participants in each band: 3;06–3;08, 3;09–3;11, 4;0–4;02, 4;03–4;05, 4;06–4;08, and 4;09–4;11.

We recruited the participants through nurseries and primary schools in the Oxford area. To minimize volunteer bias, we recruited participants using an optout protocol, which was approved by the Central University Research Ethics Committee, University of Oxford. A decile of deprivation was obtained from postcode data for each participant, where 1 represented the 10% most deprived and 10 the 10% least deprived nationally. The index of deprivation combines information for a given area, such as levels of income, employment, education, health, and crime to contribute to a single score. With the exception of the 10% most deprived, our sample included participants spanning the full range of deprivation indices. The study was explained to the participants recruited. Participants under 4 years made an oral choice regarding their participation, and those over 4 years were asked to sign an assent form. Participants were included in the sample on the basis that they had typical language abilities; had never been referred for speech and language therapy; were monolingual English speakers; and had no known intellectual, neurological, or hearing difficulties. Participants completed the Goodenough–Harris draw-a-person test (Goodenough, 1926; Harris, 1963) to ensure cognitive ability within the normal range. Although the draw-a-person test scores do not correlate well with some other measures of nonverbal reasoning (Imuta, Scarf, Pharo, & Hayne, 2013), these scores have been shown to be a predictor of later IQ in 4-year-old children (Arden, Trzaskowski, Garfield, & Plomin, 2014). In addition, we wanted a simple nonverbal measure of developmental level that could be used with children as young as 3 years and because the maturity of children's drawings shows a clear developmental trend between 3 and 5 years of age, we decided to use this as a quick play-based assessment, using the scoring methods of the draw-a-person test to quantify the maturity of drawings. Using a DSP Pure Tone Audiometer (Micro Audiometric Corporation), hearing ability was screened on the first day of assessment at three frequencies (1,000 Hz, 2,000 Hz, and 4,000 Hz) at a 25 dB level. Owing to our difficulty in locating an adequately quiet space, the results of this initial screen were inconclusive for seven participants. We further assessed these participants using the Ling six sound test, which was deemed to be an appropriate measure of the ability to hear speech at conversational loudness level.

Table 2 Relative clause test sentence examples

Relative clause type	Test sentence example
Subject intransitive	He found the girl that was hiding.
Subject transitive	He sat on the girl that was drinking the juice.
Object	The boy picked up the cup that she broke.
Oblique	The man opened the gate she jumped over.
Indirect object	She kissed the boy she poured the juice for.

Comprehension Tasks

Sentence-Verification Task

The newly devised sentence-verification task was presented on a Microsoft Surface Pro tablet device. Participants were shown 50 animations representing one of five types of relative clauses: subject (both intransitive and transitive), object, indirect object, and oblique. The 50 relative clauses, all attached to the direct object of a transitive clause, are therefore categorized as full biclausal relatives. Example test sentences are given in Table 2.

There were 10 animations for each relative clause structure, five of which matched the structure and five of which were nonmatch items. The test sentences were chosen on the basis of pilot work carried out by the first author, work completed by Diessel and Tomasello (2000, 2005), and research by Frizelle and Fletcher (2014a). 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 Inventories (Fenson, Marchman, Thal, Dale, Reznick, & Bates, 2007), to ensure an early age of acquisition. Ease of representation and pragmatic plausibility were also influencing factors.

Work on object relatives has established that not all are of equal difficulty. Friedmann, Belleti, and Rizzi (2009) demonstrated that the comprehension of these constructions is aided if the head noun phrase (NP) and the intervener—the subject of the relative clause—are of different types, for example, full NP in the case of the head and pronoun for the subject of the relative clause. In addition, work by Adani, Van Der Lely, Forgiarini, and Guasti (2010) has shown that manipulating head NP and intervener features such as number and gender (the former having a greater facilitating effect than the latter) can also aid children's comprehension of these structures. Data from Kidd et al. (2007) supported the findings reported by Friedmann et al. and also established that animacy plays a role in making object relatives easier to

comprehend. Their object relative stimuli reflected those most frequently found in input to children in having an inanimate full NP as head and a pronoun as intervener/subject of the relative clause. Accordingly, our object relative clauses reflected this earlier work in having inanimate full NPs as heads and pronominal subjects in the relative clause. In keeping with the above findings in relation to lexical restrictions and discourse relevance, we also included pronominal subjects in the oblique ("The man opened the gate *she* jumped over"), and indirect object ("She washed the baby *she* gave the bottle to") clause structures.

In natural usage, a relative clause is typically used in a contrastive fashion, for example, a sentence such as "She hugged the baby that cried" would imply the existence of another baby that did not cry. Accordingly, researchers can enhance ecological validity by structuring items so that within each animation there is an alternative to the head noun to which the relative clause refers, that is, a referent from which another can be distinguished. For example, the representation of the sentence "He laughed at the girl he threw the ball to" (where the second *he* is coreferential with the first) includes another girl in the scene who was holding a ball. Examples of test materials can be accessed at the following links: https://youtu.be/d3dz_m8zTvc; https://youtu.be/FMxYzSyCs34; https://youtu.be/rffERVi4_lc; https://youtu.be/ScPqGjoXs04; https://youtu.be/GnEvDARw7HQ. Animations were on average 6 seconds in length. As each animation began, participants were presented with a prerecorded sentence orally and asked if the animation matched the sentence that they heard. Participants were given the opportunity to hear the sentence and see the animation more than once if they needed to. However they rarely requested a repetition. They were asked to respond by touching either the smiley or sad face on the Surface Pro touch screen.

Multiple-Choice Task

The multiple-choice task was a sentence–picture matching task designed to assess the same relative clause structures as those in the sentence-verification task already described. The sentences were identical to a subset of those used in the sentence-verification task. However, the types of distractors that are feasible with static pictures in a multiple-choice task are different from those used in the sentence-verification task (see Appendixes S1 and S2 in the Supporting Information online). For example, as noted above, with only one picture to evaluate in the sentence-verification task, one can depict the restrictive reading of the relative clause (i.e., depict a baby who is not crying) whereas with the multiple-choice task, given the need to scan four different pictures, depicting

the restrictive reading would make the task overly complex and involve a large cast of characters in each set of pictures. The multiple-choice items were also part of a larger set reported on in Frizelle et al. (2017). Sentences were again prerecorded, and participants could hear the sentence more than once if they requested it, but they rarely did so. Children were given each sentence orally and were asked to choose the picture (from a set of four) that corresponded to that sentence. The other three images were distractors, which 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 (see Figure 1).

In comparison to the sentence-verification task, where participants had to choose between two (correct vs. incorrect), choosing from four pictures reduced the possibility of the participants answering correctly by chance. Therefore, in the multiple-choice task, fewer exemplars of each construction were required. In contrast to the 10 examples of each construction in the sentence-verification task, the multiple-choice task included four of each construction, yielding 20 test items in total. The pictures were presented in two formats governed by the type of verbs within the relative clause. For example, in the test sentence "She pushed the man that was reading the book," the verbs reading and pushing can occur simultaneously and can therefore be represented by four single images—the correct response and three distractors (see Figure 1). In contrast, for the sentence "The girl washed the teddy that he played with," both verbs had to be shown consecutively (the boy needs to play with the teddy before the girl washes it) and therefore required two images within each set of four (see Figure 2).

Sentence-Recall Task

The sentence-recall task was the Recalling Sentences subtest from the Clinical Evaluation of Language Fundamentals Preschool–2 UK (CELF:P–2; Wiig, Secord, & Semel, 2006). The subtest includes two practice items followed by 13 sentences, which gradually increase in length and complexity. Children are given the individual sentences orally and are required to repeat them verbatim. For the purposes of our study, the sentences were prerecorded to ensure that all participants heard them at the same rate and with the same inflection. However, it became evident in our pilot work that younger children had difficulty engaging with the task when the stimuli were not given live by the examiner. For this reason, the prerecorded sentences were not used.

Figure 2 Example of an illustration for the test sentence "The girl washed the teddy that he played with." [Color figure can be viewed at wileyonlinelibrary.com]

Procedure

A speech and language therapist with over 20 years of experience assessing child language skills administered the assessments with the majority of participants (70%). A psychologist, practicing for over 30 years and an undergraduate student of linguistics in her final year at the University of Oxford assessed the remaining 30%. Both the psychologist and the linguistics student were given training in how to administer the assessments. Participants were assessed individually in a quiet room in their respective schools. The assessments were administered in two sessions between 3 and 20 days apart. The hearing screening was carried out on the first occasion that each participant was seen. The multiple-choice and sentence-verification comprehension tasks were never administered in the same session. The sentence-recall task was audio recorded and sentences from 20% of the participants were retranscribed to check transcription reliability. Interrater transcription agreement was at 98.4%. The sequence of test sentences 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. As a validation check, assessment order was entered as a variable into a mixed-effects logistic regression and was

shown to have no significant effect, so it was dropped from the final model. For the sentence-verification task, participants simultaneously watched an animation and listened to a sentence. They were required to indicate if the animation represented the sentence that they had heard by touching the smiley face/sad face icons on the tablet screen. For the multiple-choice task, participants listened to a sentence and were asked to point to the picture that corresponded to that sentence. Verbal positive feedback, such as "well done, you are doing a great job" was given every four to five trials. Because the sentence-verification task was significantly longer than the multiple-choice task, participants received positive feedback visually (using star animations) every 10 trials. Positive feedback given for both tasks was independent of participants' performance on the tasks.

Data Analysis

Scoring for the sentence-verification task was automated, and the results were collated and stored within the application. The researcher scored the multiple-choice task in real time as the participants completed it. The initial scoring system for both tasks was binary, 1 for a correct response and 0 for an incorrect response. The scoring for the sentence-recall task was based on the CELF:P–2 scoring system (0, 1, or 2). Sentences from 20% of the participants were rescored for reliability. Interrater agreement was at 99.2%.

To allow both assessments to be compared based on similar scales, a coding system was developed that took into account the probability of obtaining a correct response by chance. The binomial theorem was used to compute the probability of obtaining a given number of items correct by chance guessing, where chance was .25 for the multiple-choice test and .50 for the sentence-verification task. A two-tiered coding system was applied; participants were given 2 if they answered four out of four items correctly on the multiple-choice task ($p = .004$) and 9 or 10 out of 10 items correctly on the sentence-verification task ($p = .01$). In the more lenient coding, participants were given 1 if they answered three out of four items correctly on the multiple-choice task and 8 out of 10 items on the sentence-verification task. With this coding, the binomial probability of a score of 1 or 2 on the multiple-choice test was .051, and the probability of a score of 1 or 2 on the sentence-verification task was .055. Thus, in both tasks, a score of 2 could be regarded as evidence that participants completely understood the construction and a score of 1 that they were close to mastery. Two or fewer correct responses out of four items on the multiple-choice test or 7 or fewer correct responses out of 10 sentences on the

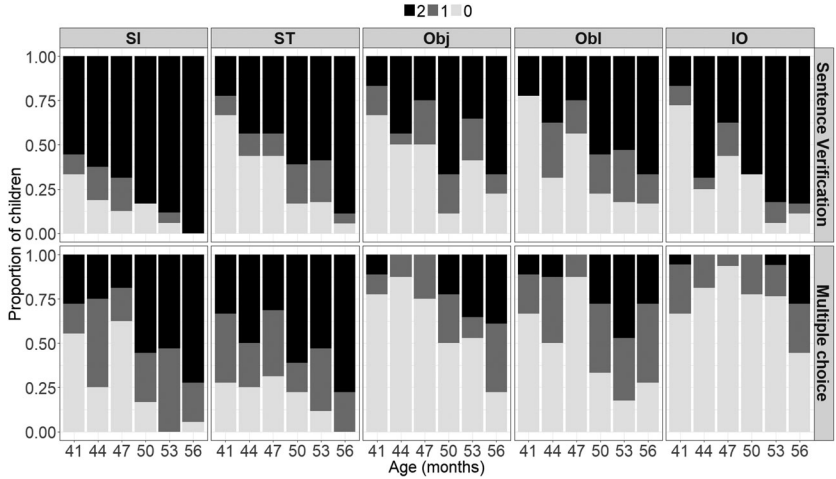

Figure 3 The proportion of participants categorized as 2 (mastery), 1 (above chance), or 0 (not mastered) for each construction, at a given age, on the sentence-verification and multiple-choice tasks. SI = subject intransitive, ST = subject transitive, Obj = object, Obl = oblique, IO = indirect object.

sentence-verification task were coded as 0. This coding allowed us to compare participants' performance across different relative clause types in both tests using a similar metric.

Results

Overall Findings

Figure 3 shows the proportion of participants categorized as 2 (mastery), 1 (above chance), or 0 (not mastered) for each construction, at a given age, on the sentence-verification and multiple-choice tasks. An examination of Figure 3 shows considerable growth in participants' comprehension of these constructions from 3 years 6 months to 4 years 11 months. This is particularly evident for the sentence-verification task where (with the exception of the subject transitive clauses) more evidence of full mastery on all constructions was demonstrated. In the three younger age bands, where participants were aged between 3 years 6 months and 4 years 2 months, the differences on full mastery between the two tasks was particularly evident. For example, younger participants showed very limited mastery of object and indirect object relative clauses on the multiple-choice task, yet a considerable percentage of participants showed full mastery on the sentence-verification task.

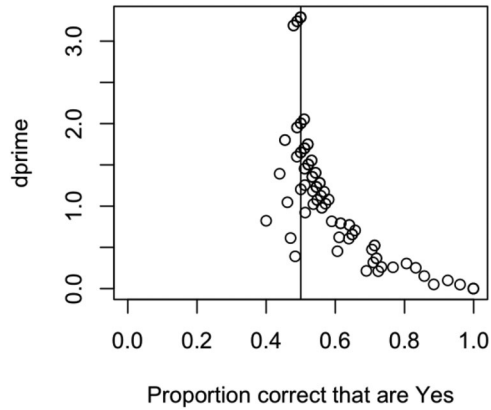

Figure 4 Plot of d' against the proportion of correct yes responses on sentence-verification task.

On the sentence-verification task, there was the possibility of response bias, whereby participants might prefer to always give a yes or no response when they failed to understand. Figure 4 shows relevant data obtained by converting the proportions of hits and false positives to a d' measure, which is plotted against the proportion of correct responses that were yes. This shows clearly that participants were biased toward a yes response when they did not understand, but for levels of d' of 1 or greater, there is a more even distribution of yes and no responses. The Pearson correlation between the d' measure and the total percentage of items correct was .87 ($df = 101$; 95% confidence interval [.81, .91]).

Sentence Verification Versus Multiple Choice

Our first research question was whether children show greater understanding of relative clauses when assessed using a sentence-verification animation task versus a multiple-choice task. To analyze the data statistically, we grouped together categories 1 and 2 for contrast with category 0. This avoided floor effects, which would have reduced sensitivity to distinguishing between the more difficult constructions (see Figure 3). Using the statistical software R (R Core Team, 2018) with lmer4 (Bates et al., 2015) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017) packages, we then fitted a mixed-effects logistic regression model to the data, which converged without error. The model was composed of main effects for task, age in months (centered), clause, and clause × task and age in months (centered) × task interactions. A random intercept for individual

Table 3 Comparison of fit for model with and without random slopes

Model	*df*	AIC	BIC	Loglik	Deviance	$\chi^2(14)$
Reduced model	13	1016.9	1081	–495.43	990.87	
Random slopes model	27	1042.7	1176	–494.34	988.68	2.19

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion.

was included into the model to quantify individual variation in baseline. The initial model was fitted without the random slope term for clause per individual, but in response to a reviewer suggestion, we subsequently added the random slope (R code can be found at https://osf.io/27gwb). The full model syntax for a generalized linear model in R using lme4 is: glmer (resp ~ ttt + Age_c + clause + clause*ttt + Age_c*ttt + (1 + clause | ID), family = binomial). A likelihood ratio test was not significant, indicating no additional benefit of the random slope structure. Furthermore, the reduced model (no random slope) had substantially lower Akaike information criterion and Bayesian information criterion values (see Table 3). In addition, no differences in the significance or magnitude of main effect were observed with or without the random slope (Table 4).

Odds ratios were calculated as measures of effect size. This represents the odds of scoring 1 versus scoring 0 given a particular condition, compared to the odds of 1 versus 0 occurring in the absence of that condition. An odds ratio of 1 indicates no difference between conditions. Table 4 shows that an effect of test, with participants performing better on the sentence-verification task than on the multiple-choice task: The odds of achieving higher scores on the sentence-verification task were 3.6 times greater than on the multiple-choice task. To evaluate the effect of the specific constructions, the intransitive relative clause score was the reference against which the other relative clause types were measured. In relation to the intransitive relatives, the odds of achieving a higher score on the oblique relatives were 1 in 4 (0.27), on the object relatives 1 in 10 (0.10), and on the indirect object relatives 1 in 20 (0.05). The model also revealed an effect of age, showing that older participants were more likely to achieve higher score on the tasks.

Although we found considerable differences between the two assessment methods at the level of full mastery, particularly with the younger participants, when we analyzed the data according to above-chance performance on the tests (2 or 1 vs. 0), there was no interaction between the method of testing and age overall. We also considered the pattern of results for full mastery (2 vs.

Table 4 Estimates of fixed effects for mixed-effects logistic regression

	Without random slope			With random slope		
Parameters	OR	95% CI	*p*	OR	95% CI	*p*
(Intercept)	4.44	[2.39, 8.25]	<.001	4.44	[2.54, 8.82]	<.001
Test type (sentence verification)	3.60	[1.49, 8.67]	.004	3.60	[1.55, 9.03]	.004
Age (centered)	2.72	[1.84,4.01]	<.001	2.72	[1.76, 3.85]	<.001
Transitive	1.93	[0.87, 4.31]	.107	1.93	[1.00, 5.00]	.101
Indirect object	0.05	[0.02, 0.11]	<.001	0.05	[0.02, 0.11]	<.001
Object	0.12	[0.05, 0.24]	<.001	0.12	[0.05, 0.23]	<.001
Oblique	0.27	[0.13, 0.56]	<.001	0.27	[0.12, 0.53]	<.001
Test type × Transitive	0.10	[0.03, 0.34]	<.001	0.10	[0.03, 0.30]	<.001
Test type × Indirect object	4.14	[1.31,13.20]	.016	4.14	[1.12, 11.32]	.041
Test type × Object	1.02	[0.33, 3.13]	.977	1.02	[0.32, 3.06]	.994
Test type × Oblique	0.53	[0.17, 1.63]	.269	0.53	[0.17, 1.55]	.243
Test type × Age	1.26	[0.87, 1.80]	.224	1.26	[0.88, 1.82]	.213

Note. OR = odds ratio. Intransitive subject relative clause is treated as reference category.

1 or 0), and the pattern did not change. However, the significant odds ratios for interaction terms with transitive subject and indirect object relatives indicated that participants' performance on these constructions differed depending on which method of assessment was used whereas this was not seen for the other constructions (see Table 4 for the fixed effects). This suggests that some constructions are more vulnerable to testing method than others.

Hierarchy of Relative Clause Structures

Our second question was whether the order of difficulty of constructions was similar on the two types of test. The interaction between test type and some clause types seen in Table 4 suggested that this was not the case. To investigate this more fully, we conducted two logistic regressions on item-level data for each test, with age in months (centered) and clause type as factors and subjects as random effects to allow for the repeated measures at item level (see Appendix S2 in the Supporting Information online), with a Tukey test (Holm adjusted *p* values) to compare clause types. Intransitive subject relative clause was the reference category. For the sentence-verification task, intransitives were

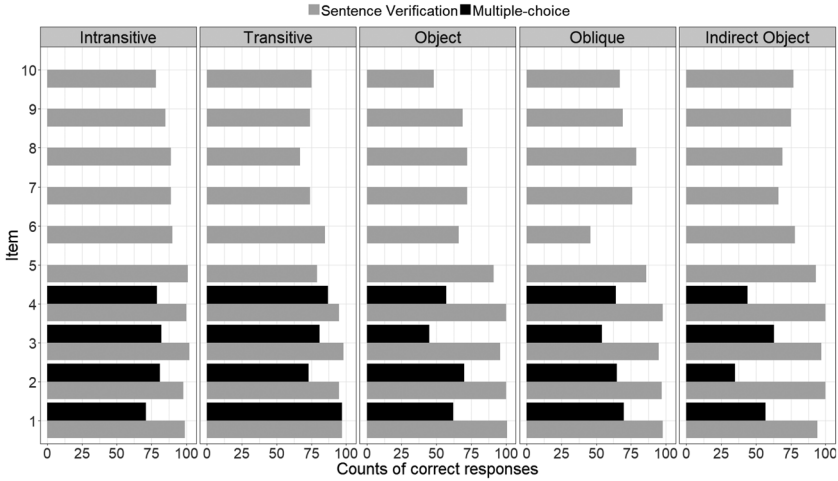

Figure 5 Bar plot of item responses for each relative clause type on the sentence-verification and multiple-choice tasks.

easier than all other clause types, but the other clause types did not differ in difficulty whereas there were three difficulty levels on the multiple-choice task. In summary, on the sentence-verification task, the pattern was intransitive subject > indirect object = transitive subject = object = oblique relatives (where > refers to "significantly greater than" and = refers to "no significant differences"). In contrast, on the multiple-choice task the ordering was transitive subject = intransitive subject > oblique = object > indirect object relatives.

Because only a subset of the items included in the sentence-verification task were included in the multiple-choice relative clause task, we were interested to see if participants showed a higher performance on some items than others—in particular, if participants performed better on the items only included in the sentence-verification task. The bar plot in Figure 5 shows that this was not the case. On the contrary, the items that were included in both tasks appeared to be easier than those that were included in the sentence-verification task only. This may reflect restrictions on the verb types that can be adequately represented in a still image. For example, verbs such as carry and hold are difficult to distinguish from each other without a representation of movement. Perhaps those dynamic verbs that can be represented in a still image are even easier to interpret in an animated form because the beginning and end of the action is represented. We were also interested in whether there were specific item effects within a particular relative clause category on either task. In relation

Table 5 Pearson correlations between variables

Variable	1	2	3	4	5
1 Sentence verification	–				
2 Multiple-choice	.56*	–			
3 Sentence recall[a]	.40*	.50*	–		
4 Age	.48*	.50*	.03	–	
5 Draw-a-person	.46*	.45*	.16	.51*	–

Note. [a]Age-scaled scores. *$p < .001$.

to the sentence-verification task, the sentences "The man opened the gate she jumped over" and "The woman took the scarf he wore" appeared to be particularly difficult (correct responses were less than 50%). In the bar plot, these correspond to object relative Number10 and oblique relative Number 6, respectively. In relation to the former, it appeared that the gates depicted in the animation were not far enough apart to be clearly interpreted as two gates, particularly by some of the younger participants. This example is shown in the YouTube links previously given. In relation to the latter, it is difficult to speculate why this item was particularly problematic. Inevitably, there will be aspects of different animations and pictures used which will affect test takers' performance, and it is not always possible to say why. This applies to the two multiple-choice task sentences with which participants had particular difficulty: "The woman washed the car he was driving" and "He smiled at the woman he cooked dinner for." These correspond to object relative Number 3 and indirect object relative Number 2, respectively.

Relationship to Sentence Recall

Our third question asked whether participants' performance on either measure of syntactic comprehension would correlate with their performance on the sentence-recall task from the CELF:P–2 (Semel et al., 2006). Table 5 provides a summary of the correlations between measures. The lack of correlation between sentence recall and age arose because these were age-scaled scores. Draw-a-person scores were missing for three participants, who were therefore excluded from the analysis. Correlations between both assessments of syntax and sentence recall scores were substantial using Davis's (1971) criteria for interpreting the magnitude of correlation coefficients. Although not as strong, associations between both measures of syntax and draw-a-person scores were also highly significant and were classified as moderate associations.

Table 6 Hierarchical linear regression: Predicting sentence recall score from sentence-verification and multiple-choice tasks, age, and draw-a-person test

Model	Predictor variables	R^2	ΔR^2	p^a
Null	Age	.001		
1	Age & draw-a-person	.028	.002	.047
2	Age, draw-a-person, & multiple-choice	.314	.286	<.001
3	Age, draw-a-person, multiple-choice, & sentence verification	.363	.049	.008

Note. Outcome variable = sentence recall. [a]Based on likelihood ratio test between subsequent models.

Given that sentence recall is considered a reflection of an individual's underlying grammatical competence, we were also interested in investigating the independent contributions of the sentence-verification and multiple-choice tasks to sentence-recall scores. Likelihood ratio tests were used to perform a hierarchical linear regression analysis, and results are shown in Table 6. We started with variables that would be expected to affect performance but were not of particular theoretical interest for this study and therefore would act as confounds. In the first step, age was entered into the model in order to control for its effect on participants' sentence-recall performance, and this was followed by draw-a-person scores. Given that the multiple-choice task is the typical format used to assess children's receptive knowledge, multiple-choice scores were entered into the model next. Multiple-choice scores explained an additional 28.6% of the variance ($p < .001$) in sentence-recall performance. Finally, sentence-verification scores were entered into the model and accounted for 5% of the variance in sentence recall over and above the multiple-choice task. The final model therefore showed that, while there was considerable overlap in the variance explained by each test, independent variance in sentence-recall ability was explained by the sentence-verification task. When the order of entry of the sentence-comprehension tests was reversed at Steps 2 and 3, then sentence verification explained 16.7% of variance at Step 2, with multiple choice then accounting for the remaining 16.8% of variance at Step 3.

Discussion

This study aimed to examine the extent to which assessment method affects assessment outcome in relation to children's understanding of complex sentences. We compared two methods of assessment: a multiple-choice

picture matching sentence comprehension task and a newly devised truth–value-judgment/sentence-verification task that was designed to optimize children's performance by using sentences presented in a pragmatically appropriate context. As we predicted, the results showed that participants performed better on the truth–value/sentence-verification task than on the multiple-choice task, and this was across the full age range. In addition, we aimed to explore whether the two assessment methods would reveal a similar order of difficulty in relation to understanding specific syntactic constructions. We predicted the same order for both assessment methods, but our results showed a different hierarchy depending on the method used. Finally, we were interested to know if results from either test would predict children's performance on an independent measure of language (sentence-recall ability). As expected, our results showed that the sentence-verification task predicted participants' sentence-recall performance, over and above that explained by the multiple-choice test. We interpret this as a validation of the sentence-verification task as a measure of syntactic knowledge.

Differences in Task Demands

As we previously discussed, a recent study by Frizelle et al. (2017) highlighted the fact that the multiple-choice picture-matching assessment method may be problematic for assessing comprehension in that it tests skills beyond those of linguistic competence. Frizelle et al. compared this method of assessing comprehension with a sentence-recall task, which (once beyond span) is considered to tap into both comprehension and production of language. They found that children could repeat sentences that they did not understand and that there was little agreement between the two measures in relation to individual differences. Although the lack of agreement between the two measures is unsettling, sentence-recall and multiple-choice methods do not purport to measure identical linguistic skills. The two methods used in this study, however, have been designed specifically to assess sentence comprehension without any additional expressive output required. If these assessment methods are equivalent, neither should compromise access to the knowledge being tested. By the same reasoning, neither assessment method should overrepresent children's abilities. Our results show that our participants across the full age range performed better on the sentence-verification task than on the multiple-choice task, causing us to question what was influencing their performance on the two measures. We suggest that differences may be driven by task design. The influence of task design on children's performance is not specific to language comprehension and has also been a topic of discussion in the field of statistical learning,

where measures have yielded conflicting results (Isbilen, McCauley, Kidd, & Christiansen, 2017).

In relation to this study, in attempting to process a sentence for syntactic comprehension, children typically form a semantic representation of the sentence while using the syntactic structure to help them assign thematic roles to the appropriate words. In the case of the sentence-verification task, children hear the target sentence and see the animation at the same time. We expect that children use typical thematic role-to-verb argument mapping to process the given sentence in real time, as they would in natural discourse. The added requirement is that they must decide whether the linguistic structure describes accurately what happened in the animation. In addition, in the sentence-verification task, children are not required to infer the same amount of information as they are in a still image. The analysis using d' gave further insights into how children approach the sentence-verification task—participants seldom rejected a correct sentence but often accepted an incorrect distractor. A tendency to respond yes might be interpreted as indicating prosociality, that is, children are biased to agree with the adult examiner. However, we suggest another interpretation, which is that children will accept a sentence that provides a good enough match to a picture. In everyday contexts, young children must continually be confronted with complex language that overwhelms their ability to interpret referents and assign thematic roles in real time; rather than simply ignoring the input, they will presumably construct meaning from partial elements that they have processed to generate a simplified and perhaps underspecified representation of meaning. Thus, we suggest that comprehension may develop from a rather general, fuzzy, and imprecise representation of meaning—which could be compatible with a range of pictured scenes—to a crisper, specific representation.

The multiple-choice task, in contrast, is presented using still images, which require participants to infer temporal relationships and movement of characters, which are often depicted by more subtle means. The need for a detailed, specified meaning representation is explicit because along with the correct answer, there are three additional distractors, requiring participants to rule out three competing alternative mappings. This results in a number of possibilities regarding how the thematic roles should be assigned and increases the processing load of the task considerably. The presence of three distractors also requires participants to store in memory the arguments associated with each verb in order to rule them out. This memory load is heightened even further by the fact that, although the structure and thematic roles are assigned

differently, the nouns in each of the distractors are the same as those in the target sentence.

Syntactic Constructions and Their Associated Distractors

Even when we compared participants' test performance based on above-chance performance (rather than full mastery), the differences between the two methods were evident. Our results showed that some constructions were more vulnerable to testing method than others. Specifically, the testing method had a large impact on our participants' understanding of indirect object relative clauses and subject transitive relative clauses. There are several factors to take into consideration when interpreting this finding. First, we can consider the role of the specific syntactic structure under consideration. Indirect object relative clauses are those in which the relativized element is the indirect object of the relative clause, and they have what is called a stranded preposition. They are considered to be one of the more complex relative clause constructions in relation to children's production (Diessel & Tomasello, 2005; Frizelle & Fletcher, 2014a). Unlike object or oblique relatives (in which the head noun can be animate or inanimate), the head noun of the indirect object relative is always animate, which means that the thematic role assigned by the verb could be applied to either noun as in the sentence "She kissed the *boy* she poured the juice for." Previous literature has suggested that when this is the case, comprehension is more difficult (Gennari & McDonald, 2009).

A second point to consider is the role of specific distractors. For some constructions—notably subject intransitives, subject transitives, and indirect objects—the multiple-choice distractors depicted the main clause with the characters reversed (e.g., the man pushing the woman vs. the woman pushing the man). This was not the case for the distractors used in the sentence-verification task. The pattern of difficulty across all constructions (see Figure 5) does not suggest that inclusion of reversible distractors is a main determinant of task difficulty, but we cannot rule out a role for this. To explore this issue fully, one would need to devise additional animation items where the distractor differed in this regard. However, these might have been much easier if they had been presented in real time because the introductory clause "the woman pushed the man" would have allowed participants immediately to reject the sentence as not matching a picture of a man pushing a woman, that is, it would not be necessary for participants to process the full sentence.

A further issue is visual complexity of distractors. To make the multiple-choice task pragmatically appropriate, indirect object relatives were depicted

Figure 6 Multiple-choice stimulus picture and still image from the sentence-verification task. [Color figure can be viewed at wileyonlinelibrary.com]

with at least four referents in each image. For example, in the sentence "He followed the girl he gave the present to" (shown in Figure 6), the correct image showed a boy giving a girl a present and was succeeded by the boy following the same girl. There are two facets of the multiple-choice task that made the illustration particularly complex: first, the temporal aspect (the boy must be seen giving the girl the present before following her—the two actions cannot happen simultaneously) and, second, the need to provide distractors showing alternative semantic–syntactic mappings regarding who did what to whom. In contrast, a still image of the corresponding animation showed the three referents included in the animation—a boy, a girl, and an alternative girl who is also holding a present. Because the animation happened in real time, the verbs could be shown consecutively as they were enacted, and the design was not based on the need for three alternate distractors (see Figure 6).

A further consideration is that constructions vary in the type of verbs (both in main and relative clause) that are required in subject transitive relative clauses. The transitive nature of the relativized verb means that the subject of the relative clause is again usually animate ("the girl that scored the goal") and, as a result, the main clause verbs are limited to those that we enact with a person or animal, such as push, pull, chase, hit, and so on (e.g., "He pushed the girl that scored the goal"). We suggest that these types of verbs are less ambiguously depicted in an animation task than when they are shown in a still image, which requires children to engage in a higher level of inference making. In contrast, in line with discourse regularities, the head noun in our object relatives and in most of our oblique relative test items was inanimate ("The dog ate the banana she dropped" or "The girl painted the wall he pointed to"), facilitating the use of a broader range of verbs.

Syntactic Hierarchy and Pragmatic Context

Both test methods showed growth in participants' understanding of complex syntax across the age range. However, our findings in relation to a hierarchy of constructions did not show consistency between the two assessments, that is, the assessments were not sensitive to the same aspects of clause complexity. The multiple-choice task showed three levels of complexity. Participants showed the highest level of understanding of subject relative clauses (both transitive and intransitive), followed by object and oblique relative clauses and last by indirect object relative clauses. This order of difficulty more or less replicated the NP accessibility hierarchy posited by Keenan and Comrie (1977). Acquisitional work has also indicated this sequence in developmental spontaneous speech data (Diessel, 2004). In contrast, the sentence-verification task showed no differences in the difficulty level of nonsubject relative clauses. In fact, with the exception of the intransitive subject relative clauses (which participants performed best on), there were no significant differences between any of the other relative clause types. The finding that intransitive subject relative clauses were the least difficult biclausal relative clause type to process is consistent with previous literature (Diessel, 2004; Diessel & Tomasello, 2005; Frizelle & Fletcher, 2014a). However, the lack of any difference between the other relative clause types has not been previously reported and may suggest that we have been underestimating children's understanding of complex sentences.

Previous literature has highlighted the impact of manipulating the lexical features of relative clauses on children's understanding of these constructions (Adani et al., 2010; Belletti, Friedmann, Brunato, & Rizzi, 2012; Friedmann et al., 2009). Additional research has investigated the effect of altering the sentences under investigation to allow for the distributional and discourse regularities of naturalistic language input (Kidd et al., 2007). However, research on the impact of manipulating the paradigm in which researchers test children has been very limited. Assessment of comprehension can never be entirely natural. Unlike in research on expressive development, where researchers can evaluate a direct product of children's language system, with receptive language, researchers have to infer what children understand from indirect evidence. Furthermore, comprehension is implicitly contrastive—a response to a single sentence is not informative because researchers need to establish whether children differentiate between sentences with different meanings. Neither sentence-verification nor a multiple-choice sentence picture matching is a natural task, but the former is easier to integrate with materials that are linguistically focused, pragmatically scaffolded, and closely related to natural discourse. This is particularly the case when animations rather than static pictures are used.

This study showed that when such materials were used, a different picture of what children are capable of understanding emerged.

Associations With Sentence Recall

Our final research question addressed the associations between the two methods of assessing syntactic comprehension and a measure of language ability. We chose sentence-recall ability as the measure of language because there is a strong link between sentence repetition and syntactic competence (Frizelle & Fletcher, 2014a; Gallimore & Tharpe, 1981; Geers & Moog, 1978; Kidd et al., 2007; McDade et al., 1982; Polišenská et al., 2015). In current thinking, sentence repetition is not merely a task of repeating a heard series of words supported by phonological short-term memory, but it is also supported by conceptual, lexical, and syntactic representations in long-term memory (Brown & Hulme, 1995; Hulme, Maughan, & Brown 1991; Klem, Melby-Lervåg, Hagtvet, Lyster, Gustafsson, & Hulme, 2015; Potter & Lombardi, 1990, 1998; Schweickert, 1993). Recalling a sentence with comprehension involves children accessing their syntactic knowledge in long-term memory to facilitate their understanding. Both the meaning of the sentence and children's linguistic knowledge are central to their ability to reconstruct the sentence for repetition. Our finding that sentence-verification scores predicted a small but significant amount of additional variance in sentence-recall ability over and above the multiple-choice task serves to validate this as a measure of syntactic comprehension.

Limitations and Future Research

In this study, we showed that the performance of typically developing children on language-comprehension tests is heavily influenced by task demands. This is in keeping with previous findings (Frizelle et al., 2017). In addition, we highlighted the role of factors such as attention, memory, and inhibition in multiple-choice comprehension tasks. From our exploration of the demands imposed by the two assessment tasks, it could be argued that to disentangle the impact of animation (which allows for real time processing) from that of the number of distractors, an intermediary condition should be included where a truth–value judgment/sentence-verification task is presented as a still image or where a multiple-choice task is presented in an animated format. However, in practice, the former would be difficult because some of the sentence meanings cannot be depicted as static images, and the latter would create new task demands by presenting four animations consecutively or simultaneously. No task design can be regarded as a pure measure of comprehension because it

will always depend on context, but it is important to be aware of the extent to which performance may vary with task.

This study focused on complex sentences where participants were required to parse different clauses and where working memory was therefore relevant to their ability to understand them (Frizelle & Fletcher, 2014b; Martin & McElree, 2009). It would be interesting to explore if assessment task demands influence children's performance on simple sentences to a similar degree. In addition, if the processing demands of multiple-choice tasks influence the performance of typically developing children, an important focus of future research would be to establish the degree to which processing demands impact on children with impaired language such as those with developmental language disorder. Because many children with language disorders have difficulties with executive functions such as inhibition, memory, and sustained attention (Archibald & Gathercole, 2006; Finneran, Francis, & Leonard, 2009; Im-Bolter, Johnson, & Pascual-Leone, 2006; Montgomery & Evans, 2009), these tasks may disadvantage them and underestimate their comprehension abilities. More generally, our study cautions against relying on any single source of data when devising accounts of children's comprehension development: Theories will be robust if they take into account converging evidence from different methods.

Conclusion

In conclusion, we compared two comprehension assessment methods and found that by varying the assessment context, we altered the results in relation to participants' understanding of complex sentences. The numerous differences between methods—in terms of distractors as well as depictions of targets—make it impossible to attribute the difference to any one factor. However, our goal was not so much to compare two formats or to pin difficulty down to any one linguistic or nonlinguistic feature as to consider whether by moving away from a more traditional multiple-choice format for complex sentence understanding we could create an assessment method that would enable children to reveal more linguistic knowledge. Our findings suggest that when sentences are presented in a context that better reflects how language is processed in natural discourse, young children have a greater understanding of complex sentences than would be indicated by multiple-choice testing. Multiple-choice tests may be useful when researchers want to see how robust comprehension is when children are forced to focus only on grammatical structure and have to ignore other information that may be competing or distracting, but interpretation needs to take into account the role of other factors such as attention, memory, and

inhibition. Thus, in addition to testing complex sentences, researchers need other items that pose similar task demands but that use simpler sentences. In addition, in a well-designed multiple-choice test, analysis of error patterns can help distinguish genuine grammatical difficulties from more general cognitive impairments (Bishop, 2003). We conclude that those designing and administering language assessments need to critically appraise the tools that they are using and to reflect on whether a test is actually measuring what it claims to measure. The current technological era opens the way to developing assessment tools that allow testing of children's understanding in real time. Our results have implications for those who work clinically as well as for researchers who are investigating children's language comprehension ability.

Final revised version accepted 29 September 2018

References

Adams, C., Coke, R., Crutchley, A., Hesketh, A., & Reeves, D. (2001). *Assessment of Comprehension and Expression 6–11* [Measurement instrument]. London, England: NFER-Nelson. https://doi.org/10.1037/t05199-000

Adani, F., Van Der Lely, H. K. J., Forgiarini, M., & Guasti, M. T. (2010). Grammatical feature dissimilarities make relative clauses easier: A comprehension study with Italian children. *Lingua*, *120*, 2148–2166. https://doi.org/10.1016/j.lingua.2010.03.018

Archibald, L., & Gathercole, S. (2006). Short term and working memory in specific language impairment. *International Journal of Language & Communication Disorders*, *41*, 675–693. https://doi.org/10.1080/13682820500442602

Arden, R., Trzaskowski, M., Garfield, V., & Plomin, R. (2014). Genes influence young children's human figure drawings and their association with intelligence a decade later. *Psychological Science*, *25*, 184–350. https://doi.org/10.1177/0956797614540686

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://www.jstatsoft.org/article/view/v067i01/0

Belletti, A., Friedmann, N., Brunato, D., & Rizzi, L. (2012). Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua*, *122*, 1053–1069. https://doi.org/10.1016/j.lingua.2012.02.007

Bishop, D. V. M. (2003). *Test for Reception of Grammar* [Measurement instrument]. London, England: Pearson.

Brown, G. D. A., & Hulme, C. (1995). Modelling item length effects in memory span: No rehearsal needed? *Journal of Memory and Language*, *34*, 594–621. https://doi.org/10.1006/jmla.1995.1027

Clark, E. V. (1971). On the acquisition of the meaning of *before* and *after*. *Journal of Verbal Learning and Verbal Behavior*, *10*, 266–275. https://doi.org/10.1016/S0022-5371(71)80054-3

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. (1985). The acquisition of structural restrictions on anaphora. In S. Berman, J. W. Choe, & J. McDonagh (Eds.), *Proceedings of NELS 16*. Amherst, MA: Graduate Linguistic Student Association of the University of Massachusetts.

Crain, S., Thornton, R., Boster, C., & Conway, L. (1996). Quantification without qualification. *Language*, *5*, 83–153. https://doi.org/10.1207/s15327817la0502_2

Davis, J. A. (1971). *Elementary survey analysis*. Englewood Cliffs, NJ: Prentice-Hall. https://doi.org/10.2307/2577148

de Villiers, J. G., Tager-Flusberg, H., Hakuta, K., & Cohen, M. (1979). Children's comprehension of relative clauses. *Journal of Psycholinguistic Research*, *8*, 499–518.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/cbo9780511486531

Diessel, H., & Tomasello, M. (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, *11*, 131–151. https://doi.org/10.1515/cogl.2001.006

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, *81*, 1–25. https://doi.org/10.1353/lan.2005.0169

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories* [Measurement instrument]. Baltimore, MD: Brookes.

Finneran, D., Francis, A., & Leonard, L. (2009). Sustained attention in children with specific language impairment. *Journal of Speech and Hearing Research*, *52*, 915–929. https://doi.org/10.1044/1092-4388(2009/07-0053)

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of Verbal Learning and Verbal Behavior*, *2*, 121–135. https://doi.org/10.1016/s0022-5371(63)80076-6

Friedmann, N., Belletti, A., & Rizzi, L. (2009). Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua*, *119*, 67–88. https://doi.org/10.1016/j.lingua.2008.09.002

Frizelle, P., & Fletcher, P. (2014a). Relative clause constructions in children with specific language impairment. *International Journal of Language & Communication Disorders*, *49*, 255–264. https://doi.org/10.1111/1460-6984.12070

Frizelle, P., & Fletcher, P. (2014b). The role of memory in processing relative clauses in children with specific language impairment. *American Journal of Speech-Language Pathology*, *24*, 47–59. https://doi.org/10.1044/2014_ajslp-13-0153

Frizelle, P., O'Neill, C., & Bishop, D. V. M. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language*, *44*, 1–23. https://doi.org/10.1017/S0305000916000635

Gallimore, R., & Tharp, R. G. (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning*, *31*, 369–392. https://doi.org/10.1111/j.1467-1770.1981.tb01390.x

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*, 380–391. https://doi.org/10.1044/jshd.4303.380

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*, 1–23. https://doi.org/10.1016/j.cognition.2008.12.006

Gerken, L., & Shady, M. E. (1996). The picture selection task. In D. McDaniel, C. McKee, & H. Smith Cairns (Eds.), *Methods for assessing children's syntax* (pp. 55–76). Cambridge, MA: MIT Press.

Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. New York, NY: World Books.

Goodluck, H., & Tavakolian, S. (1982). Competence and processing in children's grammar of relative clauses. *Cognition*, *11*, 1–27. https://doi.org/10.1016/0010-0277(82)90002-6

Gordon, P. (1996). The truth-value judgment task. In D. McDaniel, C. McKee, & H. Smith-Cairns (Eds.), *Methods for assessing children's syntax* (pp. 211–231). Cambridge, MA: MIT Press.

Gordon, P., & Chafetz, J. (1986, October). *Lexical learning and generalization in the passive acquisition*. Paper presented at the 11th Annual Boston University Conference on Language Development, Boston, MA.

Hamburger, H., & Crain, S. (1982). Relative acquisition. *Language Development: Syntax and Semantics*, *1*, 245–274.

Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity*. New York, NY: Harcourt, Brace, Jovanovich.

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*, 685–701. https://doi.org/10.1016/0749-596x(91)90032-f

Im-Bolter, N., Johnson, J., & Pascual-Leone, J. (2006). Processing limitations in children with specific language impairment: The role of executive function. *Child Development*, *77*, 1822–1841. https://doi.org/10.1111/j.1467-8624.2006.00976.x

Imuta, K., Scarf, D., Pharo, H., & Hayne, H. (2013). Drawing a close to the use of human figure drawings as a projective measure of intelligence. *PloS ONE*, *8*, e58991. https://doi.org/10.1371/journal.pone.0058991

Isbilen, E., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. In G. Gunzelmann, A. Howes, T. Tebrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 564–569). Austin, TX: Cognitive Science Society.

Jisa, H., & Kern, S. (1998). Relative clauses in French children's narrative texts. *Journal of Child Language*, *25*, 623–652. https://doi.org/10.1017/s0305000998003523

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. In *Linguistic Inquiry*, *8*, 63–99.

Kidd, E., & Bavin, E. L. (2002). English-speaking children's comprehension of relative clauses: Evidence for general-cognitive and language-specific constraints on development. *Journal of Psycholinguistic Research*, *31*, 599–617. https://doi.org/10.1023/a:1021265021141

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*, 860–897. https://doi.org/10.1080/01690960601155284

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*, 146–154. https://doi.org/10.1111/desc.12202

Kuznetsova, A, Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://www.jstatsoft.org/article/view/v082i13

Limber, J. (1976). Unraveling competence, performance and pragmatics in the speech of young children. *Journal of Child Language*, *3*, 309–318. https://doi.org/10.1017/s0305000900007200

Martin, A. E., & McElree, B. (2009). Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*, 1231–1239. https://doi.org/10.1037/a0016271

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*, 19–24. https://doi.org/10.1044/jshd.4701.19

McDaniel, D., & McKee, C. (1998). *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.

McGhee, R. L, Ehrler, D. J., & DiSimoni, F. (2007). *Token Test for Children–Second Edition* [Measure instrument]. Austin, TX: Pro-Ed.

Montgomery, J., & Evans, J. (2009). Complex sentence comprehension and working memory in children with specific language impairment. *Journal of Speech,*

Language, and Hearing Research, *52*, 269–288. https://doi.org/10.1044/1092-4388(2008/07-0116)

Philip, W. (1991). Spreading in the acquisition of universal quantifiers. *West Coast Conference on Formal Linguistics*, *10*, 359–373.

Philip, W. (1992). Distributivity and logical form in the emergence of universal quantification. *Semantics and Linguistic Theory*, *2*, 327–346. https://doi.org/10.3765/salt.v2i0.3029

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language & Communication Disorders*, *50*, 106–118. https://doi.org/10.1111/1460-6984.12126

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*, 633–654. https://doi.org/10.1016/0749-596x(90)90042-x

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, *38*, 265–282. https://doi.org/10.1006/jmla.1997.2546

R Core Team (2018). *R: A Language and Environment for Statistical Computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Roeper, T., & De Villiers, J. (1993). The emergence of bound variable structures. In E. Reuland & W. Abraham (Eds.), *Knowledge and language* (pp. 105–139). Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-011-1840-8_6

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory and Cognition*, *21*, 168–175. https://doi.org/10.3758/bf03202729

Semel, E., Wiig, E. M., & Secord, W. (2006). *Clinical Evaluation of Language Fundamentals–Fourth Edition UK* [Measurement instrument]. London, England: 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*, 272–281. http://doi.org/10.1016/S0022-5371(74)80064-2

Slobin, D. I., & Welsh, C. A. (1968). Elicited Imitation as a research tool in developmental psycholinguistics. *Working Papers of the Language Behavior Research Laboratory*, *10*, 1–21.

Tavakolian, S. (1977). *Structural principles in the acquisition of complex sentences* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, MA.

Wiig, E. H., Secord, W. A., & Semel, E. (2006*). Clinical Evaluation of Language Fundamentals–Preschool 2 UK* [Measurement instrument]. London, England: Pearson Assessment. https://doi.org/10.1037/e313642005-013

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Distractor List for the Sentence-Verification Task.
Appendix S2. Distractor List for the Multiple-Choice Task.
Appendix S3. Correlations Between Scores for Each Relative Clause Type.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Testing Children's Understanding of Language: How Well They Do Depends on How We Test It

What This Research Was About And Why It Is Important

The focus of this study was on children's understanding of complex sentences. Complex sentences are made up of two or more simple sentences. For example, the complex sentence "He found the girl that was hiding" includes two simpler sentences "He found the girl" and "The girl was hiding." Children often do poorly on tests designed to assess their understanding of complex sentences. In this study, the researchers examined the effect of two testing methods on how typically developing children understand a specific type of complex sentence—a sentence containing a relative clause (e.g., "She tickled the boy that laughed" and "She pointed to the man that was reading"). The researchers found that children's understanding was heavily influenced by the testing method used. This is important because it suggests that some tests that are widely used in clinical and educational settings may underestimate children's understanding.

What the Researchers Did

- The researchers tested 103 children between the ages of 3 years 6 months and 5 years. All children had typical language abilities, had never been referred for speech and language therapy, were monolingual English speakers, and had no known intellectual, neurological, or hearing difficulties.
- Children were tested on their understanding of identical sentences, using two methods:
  - One testing method was a multiple-choice task, in which children heard a sentence and then selected a picture, from among four *still* images, that best corresponded to the sentence. For instance, after hearing "She pushed the man that was reading the book," children selected the image of the woman pushing a man holding a book.

  - The other testing method was an animated task, in which the child heard a sentence while simultaneously being shown an *animation* depicting an action. The child was required to decide if the sentence described the animation (e.g., https://youtu.be/d3dz_m8zTvc and https://youtu.be/GnEvDARw7HQ).
- The scoring of children's performance was adjusted to account for the fact that in the multiple-choice method the child had a 25% chance of choosing the correct picture compared to a 50% chance of making the correct choice in the animated task.

What the Researchers Found

- Children could understand a greater number of sentences when the animated testing method was used compared to the multiple-choice method.
- The same sentence presented children with different degrees of difficulty depending on the testing method.
- The order of difficulty of the sentences (most to least difficult to understand) varied between the test methods.

Things to Consider

- The performance of typically developing children on tests assessing their understanding of language is heavily influenced by how children are tested.
- The most commonly used clinical testing method, which relies on multiple-choice testing of children's understanding of language, may underestimate children's ability to understand complex sentences.
- Other factors (beyond language), such as attention and memory, likely influence scores from multiple-choice tests. For example, attending to, comparing, and selecting among multiple alternatives might overburden children's memory.
- The role of attention and memory factors in multiple-choice tests is particularly relevant for children with developmental language disorders, many of whom have difficulties with memory and attention in addition to language. Therefore, these children may be disadvantaged when tested using multiple-choice tests.
- Researchers and practitioners should be cautious when describing the development of children's understanding of language using results obtained from a single testing method.

How to cite this summary: Frizelle, P., Thompson, P. A., Duta, M., & Bishop, D. V. M. (2019). Testing children's understanding of language: How well they do depends on how we test it. *OASIS Summary* of Frizelle et al. in *Language Learning*. https://oasis-database.org