LANGUAGE LEARNING AND DEVELOPMENT
2020, VOL. 16, NO. 4, 426–450
https://doi.org/10.1080/15475441.2020.1820340

Routledge
Taylor & Francis Group

Statistical Learning in the Visuomotor Domain and Its Relation to Grammatical Proficiency in Children with and without Developmental Language Disorder: A Conceptual Replication and Meta-Analysis

Imme Lammertink[a], Paul Boersma[a], Frank Wijnen[b], and Judith Rispens[a]

[a]Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands; [b]Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Children with Developmental Language Disorder (DLD) have difficulties acquiring the grammatical rules of their native language. It has been proposed that children's detection of sequential statistical patterns correlates with grammatical proficiency and hence that a deficit in the detection of these regularities may underlie the difficulties with grammar observed in children with DLD. Although there is some empirical evidence supporting this claim, individual studies, both in children with and without DLD, vary in the strength of their reported associations. The aim of the present study is therefore to evaluate the evidence for the proposed association. This is achieved by means of (a) a conceptual replication study on 35 children with DLD and 35 typically developing children who performed the serial reaction time task and a test of grammatical proficiency and (b) a meta-analysis over 19 unique effect sizes, which collectively examined the serial reaction time task – expressive grammar correlation in 139 children with DLD and 573 typically developing children. Both outcomes provide no evidence for or against the existence of the proposed association. Neither do they provide evidence for a difference in the strength of the association between both groups of children.

When acquiring their native language, children unconsciously detect and process structural regularities that facilitate word extraction, the induction of phonological and grammatical categories and the representation of (morpho)syntactic rules (Erickson & Thiessen, 2015; Mintz, 2003; Saffran et al., 1996; Wijnen, 2013). It has been proposed that children detect and process these regularities via *statistical learning*. Evidence that statistical learning may play a role in language development comes from two different sources. Firstly, a number of studies has reported on associations between children's statistical learning ability and different aspects of language (vocabulary size: e.g., Evans et al., 2009; syntactic processing: Kidd, 2012; Kidd & Arciuli, 2016; Misyak et al., 2010; Misyak & Christiansen, 2012; grammar: Hamrick et al., 2018; reading: Arciuli, 2018 and spelling: Treiman, 2018). Secondly, there is evidence for a statistical learning deficit in children who have Developmental Language Disorder (DLD, Evans et al., 2009; Hsu & Bishop, 2014; Lammertink et al., 2017; Lammertink, Boersma, Wijnen et al., 2020; Obeid et al., 2016). By definition, children with DLD exhibit difficulties with language, across multiple areas (lexicon, phonotactics, morphology, morphosyntax, syntax, discourse) in the absence of a known biomedical cause, intellectual disability, or unfavorable psycho-social/educational conditions (Bishop et al., 2017). Despite heterogeneity in the

CONTACT Imme Lammertink immelammertink@gmail.com Amsterdam Center for Language and Communication, The Netherlands 1012 VB, University of Amsterdam

language difficulties observed across children with DLD, almost all children with DLD struggle with the acquisition of the rule-based aspects (i.e., morphology, syntax, morphosyntax, phonology and phonotactics) of language (Leonard, 2014). Given that the detection of these rule-based aspects of language may depend on the detection of sequential statistical regularities, their problems with these components of language may be explained by a statistical learning deficit (or a procedural learning deficit, see below; Evans et al., 2009; Hsu & Bishop, 2014; Lammertink et al., 2017; Lammertink, Boersma, Wijnen et al., 2020; Obeid et al., 2016; Ullman & Pierpont, 2005; Ullman & Pullman, 2015).

A deficit in the detection of sequential regularities

The serial reaction time task (task design is explained in more detail below) is frequently used to assess children's sensitivity to sequential statistical regularities (i.e. sensitivity to differences in the transitional probability from one element to another element over time). Sensitivity to such sequential statistical information has been proposed to underlie the acquisition of grammatical rules in language. For example, in the English present tense, singular subjects frequently co-occur with [s]-marking on the verb (subject–verb agreement as in *the child* walk*s*). In order to learn subject–verb marking, children need to detect that there is a grammatical relation between a singular subject and verb–plus–[s] marking. Other than sequential statistical regularities, it has also been shown that people are sensitive to distributional statistics (e.g., Maye et al., 2002) and cross-situational statistics (Smith & Yu, 2008; Yu & Smith, 2007). However, the focus of the present study is on children's detection of sequential statistical regularities and the relation between sequential statistical learning and grammatical proficiency and we use the terms statistical learning and sequential statistical learning interchangeably in this paper.

Sensitivity to sequential regularities also plays a key role in the declarative/procedural model of language (Ullman, 2014) and the associated procedural learning deficit hypothesis (Ullman & Pierpont, 2005; Ullman & Pullman, 2015). In short, and skipping over the nuances, the declarative/procedural model of language states that the acquisition of rule-based aspects of language (such as grammar) is supported by a procedural memory system, whereas the acquisition of lexical knowledge is linked to a declarative memory system. Similar to predictions from the statistical learning literature, the declarative/procedural model of language predicts a correlation between children's sensitivity to sequential regularities and their grammatical proficiency. Furthermore, the procedural learning deficit hypothesis also predicts reduced sensitivity to sequential statistical regularities in children with DLD as compared to their typically developing peers. According to the procedural learning deficit hypothesis, this declarative learning mechanism is relatively spared in children with DLD, and children with DLD may even compensate their procedural learning deficit via declarative learning. That is, in learning the grammatical rules of their language, children with DLD may rely more on their declarative learning system than their procedural learning system (sometimes also referred to as the *declarative memory compensation hypothesis*, see Ullman & Pullman, 2015). This declarative memory compensation hypothesis predicts weaker associations between procedural learning and grammatical proficiency in children with DLD as compared to typically developing children (note that this weaker association does not mean that the hypothesis predicts *no* correlation between procedural learning and grammar proficiency in children with DLD; as explained in Lum et al., 2012, it is still likely that such an association also exists in children with DLD). To the best of our knowledge, statistical learning deficit accounts do not necessarily predict a difference in the strength of the correlation between both groups of children.

Thus, both the statistical learning deficit hypothesis and the procedural deficit hypothesis argue that children with DLD may have a deficit in their detection of sequential patterns and both accounts predict that typically developing children outperform children with DLD on any learning task that requires the detection of sequential statistical patterns (the serial reaction time task being a prime example of such a task). Evidence for the existence of a sequential learning deficit in children with DLD as compared to typically developing children comes from studies that investigated this type of learning in both groups of children in the auditory domain (see meta-analysis across 10 studies by Lammertink et al., 2017, and meta-analysis across 14 studies by; Obeid et al., 2016), the visuomotor domain (see meta-analysis across eight studies by Lum et al., 2014, and the meta-analysis by Obeid et al.) All three meta-analyses conclude that the detection of sequential regularities is, on average, not as effective in people with DLD as compared to people without DLD. Interestingly, the standardized effect size for the difference between both populations is largest in the meta-analysis on auditory verbal statistical learning (Lammertink et al., 2017). Lammertink et al. speak of a medium to large disadvantage in auditory verbal statistical learning in people with DLD. The difference between both populations is smallest in the meta-analysis on visuomotoric statistical learning (Lum et al., 2014). In this meta-analysis, the confidence interval for the standardized difference between both groups ranges from 0.072 to 0.584, meaning that we can speak of a small to medium statistical learning disadvantage for people with DLD in the visuomotor domain. Although these outcomes may suggest that the statistical learning deficit for DLD is largest in the auditory verbal domain, it should be noted that the studies included in both meta-analyses also differed in the procedures by which statistical learning is measured. The auditory verbal statistical learning tasks that were included in the meta-analysis by Lammertink et al. (2017) almost all assessed statistical learning with an offline measure (grammaticality judgments after learning took place), whereas all studies included in the meta-analysis by Lum et al. (2014) assessed statistical learning with the serial reaction time task – an online measure. As we also conclude at the end of this paper, it is likely that measurement procedures impact the size of the statistical learning deficit.

Both the statistical learning deficit hypothesis and the procedural learning deficit hypothesis predict that children's performance on the serial reaction time task correlates with grammatical proficiency. A quantitative summary (meta-analysis) of studies investigating such associations in typically developing children learning their first language provided evidence that this is indeed the case (Hamrick et al., 2018). Although the correlation between serial reaction time task performance and grammatical proficiency in children with DLD has been explored (see next section) in several studies, a quantitative summary of all these studies does not exist yet, but is needed in order to obtain an estimate of the strength of the sequential statistical learning – grammatical proficiency relationship in children with and without DLD. Such a quantitative summary (meta-analysis) on the correlation between serial reaction time performance and grammatical proficiency in children with DLD is of particular interest, because the correlation analyses reported in the individual studies are often exploratory or secondary to the DLD-typically developing group comparison on the statistical learning task. A meta-analysis with a focus on this correlation allows us to not only assess its robustness, but may help us explain variation in reported results, by exploring (potential) moderators that are difficult to assess in a single study (e.g., the effect of age and sequence type [first-order conditional versus second-order conditional], as explained later on; Black & Bergmann, 2017).

Statistical learning and grammatical proficiency: the need for replication

The discussion above reveals that there is some (albeit mostly exploratory) empirical evidence that sequential statistical learning (measured with the serial reaction time task) correlates with grammatical proficiency in typically developing children. At the same time, a closer look at the outcome of Hamrick et al.'s meta-analysis reveals that the 95% confidence interval of the average weighted correlation between serial reaction time task proficiency and grammatical proficiency ranges from +.009 to +.495. This means that, in the sense of Cohen (1992), the strength of the association in typically developing children varies between "small" and "medium to large". This relatively wide confidence interval indicates that the strength of the associations reported in individual studies varies strongly. Altogether, these results suggest a large variability in the size and existence of the proposed association between children's serial reaction time performance and their grammatical proficiency, and thus that the association may not be as robust as commonly thought.

Motivated by these large differences in observed associations, as well as the general replication crisis and the documented existence of publication biases ("file drawer problem") in psychology (e.g., Open Science Collaboration, 2015; Rosenthal, 1979), the aim of the present study is to (again) evaluate the existence and strength of the association. This is done by (a) a conceptual replication of previous experiments on a visuomotoric statistical learning deficit in children with DLD and (b) a quantitative summary (meta-analysis) of the studies that explored the proposed association between serial reaction time performance and grammatical proficiency in children with and without DLD. This meta-analysis also allows us to explore if a publication bias is likely to exist and to explore whether the serial reaction time task-grammatical proficiency correlation differs between children with and without DLD (as predicted by the declarative memory compensation hypothesis). Our meta-analysis serves a different goal than the meta-analysis on serial reaction time performance and grammatical proficiency conducted by Hamrick et al. (2018). Hamrick et al. aimed to test the predictions of the declarative/procedural model in first and second language learners (Ullman, 2014), whereas we focus (a) on the relation between serial reaction time performance and grammatical proficiency only, leaving the relationship between declarative learning and lexical knowledge aside, and (b) we focus on different populations, namely children with and without DLD. This different focus makes our analysis different from the one conducted by Hamrick et al. Our meta-analysis is also different from the meta-analysis on serial reaction time task performance by Lum et al. (2014). Lum et al. assessed the group difference in serial reaction time task performance whereas our meta-analysis assesses the strength between children' serial reaction time task performance and grammatical proficiency at the individual level. As explained above, a meta-analysis on the strength of the correlation in both groups is of particular interest, because in previous studies these correlations are often reported as exploratory analyses only.

The serial reaction time task

The serial reaction time task is one of the most commonly used tasks to assess children's sensitivity to a fixed sequence in the visuomotor domain. In this fixed sequence ("structured trials"), the appearance of a visual stimulus follows a repeating sequence of predefined positions on a computer screen. In the task, sensitivity to sequential structure is usually operationalized as the difference in response times to structured versus random trials. After repeated exposure to structured trials, random trials elicit slower responses than structured trials. After the introduction of the serial reaction time task by Nissen and Bullemer (1987), different versions of the task have been used. These versions differ, amongst other factors, in the length of the repeating sequence, in the sequence structure (first-order conditional versus second-order or higher-order conditional, explained below), in the response mode used (response box, keyboard, touch screen), and in the number of trials to which participants are exposed. These aspects may impact performance: the meta-analysis on serial reaction time performance in children with and without DLD from Lum et al. (2014), for example, showed that longer exposure to the sequenced trials leads to smaller differences in performance between children with and without DLD.

In the experimental part (i.e. our conceptual replication) of the present study, we use a serial reaction time that is based on the one used by Lum and Kidd (2012, for differences between our task and the task of Lum and Kidd, see Materials section). We decided to work with this serial reaction time task as the design of this task is comparable, in terms of the sequence type used (first-order conditional) and the block structure used (structured versus random blocks), to serial reaction time tasks that are commonly used to assess the presence of a visuomotor statistical learning deficit in children with DLD (e.g., Clark & Lum, 2017; Conti-Ramsden et al., 2015; Hsu & Bishop, 2014; Park et al., 2018). Thus, our experimental study can be seen as a conceptual replication of earlier work on the presence of a visuomotoric statistical learning deficit in children with DLD. That is, our task design does not differ in any significant way from earlier studies on this topic (for a definition of the term "conceptual replication" see Black & Bergmann, 2017).

As will also become clear from our meta-analysis, not all studies on serial reaction time task performance in children with and without DLD work with first-order conditional sequences, however. Some studies also assessed the size of the learning deficit using second-order (or even higher-order) conditional sequences. In first-order conditional sequences, each position can be predicted (albeit with varying degrees of probability) from the previous position and thus the sequence can be learned from adjacent dependencies. In second-order conditional sequences, each position occurs equally often and also each adjacent pair of positions occurs equally often; therefore, all pairwise transitions are ambiguous and the next position can only be learned from the previous two positions (Cohen et al., 1990). The use of first-order conditional sequences versus second-order conditional sequences may impact the strength of the association between serial reaction time performance and grammatical proficiency, as learning second-order conditional sequences may require different (or additional) cognitive processes than learning first-order conditional sequences (Clark et al., 2019; Wilson et al., 2018). Also, second-order conditional structure may more closely mimic the long-distance dependencies often reflected in the morphological and morphosyntactic rules of natural languages than the adjacent dependencies in first-order conditional sequences (Wilson et al., 2018). Our meta-analysis (in the second part of this paper) explores if the strength of the association between serial reaction time performance and grammatical proficiency depends on the use of second-order conditional sequences versus first-order conditional sequences.

Measures of grammatical proficiency

Tests used to assess grammatical proficiency in children vary widely on a range of parameters. Amongst other things, these tasks differ in what grammatical structures they target, and also in their assessment of children's *receptive* grammatical proficiency or *expressive* grammatical proficiency. For example, some of the studies included in our meta-analysis focused on a specific grammatical structure, such as past-tense production (e.g., Lum & Kidd, 2012; Mimeau et al., 2016) whereas others used broader tasks of grammatical proficiency, such as the sentence recall task from the Clinical Evaluation of Language Fundamentals (CELF; Semel et al., 2010) test battery (e.g., Hani, 2015; Obeid, 2017; Park et al., 2018).

In our experimental study (first part of this paper) we use the sentence recall task (a subset of the Dutch Clinical Evaluation of Language Fundamentals test battery; CELF-4-NL; Semel et al., 2010) as a measure of children's grammatical proficiency. Poor sentence recall is a robust clinical marker of DLD (Conti-Ramsden et al., 2001). In the task children are asked to recall sentences. These sentences become increasingly longer and more complex as the task continues. The task includes different sentence types (e.g., passives, declaratives, relative clause constructions) and targets different morphosyntactic processes (subject–verb agreement, past-tense production, pluralization). Performance on the sentence recall task has been shown to correlate with measures of past-tense production and grammatical usage of verbs in third person singular in both children with and without DLD (Conti-Ramsden et al., 2001). Although successful sentence recall also depends on lexical knowledge and other cognitive processes such as verbal short-term memory, verbal long-term memory (Conti-Ramsden et al., 2001) and working memory (Frizelle et al., 2017), Polišenská et al. (2015) demonstrate that "children's ability to repeat sentences is more dependent on their familiarity with morphosyntax than semantics or prosody [...]". As such we consider the sentence recall task as a good proxy of children's (expressive) grammatical proficiency.

Study 1: experimental study

Methods experimental study

Participants

Thirty-seven children with DLD and fifty-nine typically developing children, aged between seven and twelve years of age, participated in the experiment. We informed everyone involved in the recruitment process that recruitment and testing had to fit within a predetermined testing period that ran from January 2017 to March 2018, and we recruited and tested as many children as possible in the available recruitment time. We obtained approval from the ethical review committee of the University of Amsterdam, Faculty of Humanities. For the participants with DLD, their parents or caregivers gave informed consent prior to their children's participation in the study. We obtained passive informed consent from the parents or caregivers of the typically developing participants before the start of the study.

As explained in the Procedure section, the same children with and without DLD also participated in Lammertink, Boersma, Wijnen et al. (2020) and in Lammertink, Boersma, Rispens et al. (2020), but there is no overlap in the tasks. Furthermore, data from a subset of the typically developing children that participated in this study are reported on in Van Witteloostuijn, 2020).

Children with DLD

We recruited children with DLD through four national organizations in the Netherlands (The Royal Auris Group, the Royal Kentalis Group, Pento, Viertaal) and through an association for parents of children with DLD (stichting Hoormij). All children had received the diagnosis of DLD by licensed clinicians before participating in the present study, and were additionally selected to meet all of the following criteria: (a) they had scored at least 1.5 standard deviations below the norm on two out of four subscales (speech production, auditory processing, grammatical knowledge, lexical semantic knowledge) of a standardized language assessment test battery administered by a licensed clinician (but not as part of our own test battery); (b) at least one of their parents was a native speaker of Dutch; and (c) they had not been diagnosed with autism spectrum disorder, attention deficit hyperactivity disorder, or other (neuro)physiological problems. Finally, our test battery included the Raven Progressive Matrices subtest (Raven et al., 2003), a standardized measure of nonverbal intelligence, on which the participants had to obtain a percentile score of at least 17%, which is the lower bound of the normal range, to be included in our final sample. This means that the children in our sample also met the criterion for having specific language impairment (for a discussion on the labels DLD versus specific language impairment, see Bishop et al., 2017). After testing, we had to exclude two children with DLD: one child because of technical problems and one child because of a hearing problem that had only been diagnosed during the testing period.

Typically developing children

We recruited the typically developing children through four different primary schools across the Netherlands. We used the Raven Progressive Matrices subtest (Raven et al., 2003), the one-minute-real-word reading test (Brus & Voeten, 1979), the two-minute-nonce-word reading test (Van den Bos et al., 1994), a test of spelling (Braams & de Vos, 2015) and the sentence recall test from the CELF-4-NL (Semel et al., 2010) to determine if children met our inclusion criteria for the typically developing children (all these tasks were part of our own task battery, see Procedure section). We excluded children that scored below the normal range on the Raven Progressive Matrices and/or on two or more of the four language tasks mentioned above. Additionally, we also excluded children from the typically developing group if they had been diagnosed with autism spectrum disorder, attention deficit hyperactivity disorder, or with other (neuro)physiological problems. In total, we excluded five children by the first criterion and one child by the second criterion. From the remaining 53 typically developing children, we selected 35 children that matched best with our DLD sample, considering age, gender, socioeconomic status (on the basis of postal code; Sociaal en Cultureel Planbureau, 2017) and nonverbal intelligence (Raven et al., 2003). Our final sample included 35 children with DLD (7

Table 1. Overview of participants' age, gender and socioeconomic status as well as the mean raw and, when available, standardized scores of the children with and without DLD on the raven progressive matrices task, the one-minute-real-word reading task, the two-minute-nonce-word reading task, a test of spelling and the sentence recall task (measure of morphosyntactic knowledge).

		DLD	TD	Comparison between DLD and TD		
		N = 35 (F = 7,M = 28)	*N*= 36 (F = 9,M = 26)	*t*	*p*	95% CI
Age (years;months)						
Raw	Mean	9;1	9;1	0.028	0.98	−0.3 to +0.3
	Range	7;8 to 10;4	7;8 to 10;4			
Socioeconomic status (SES; based on postal code; Sociaal en Cultureel Planbureau, 2017)						
Raw	Mean	+0.23	−0.030	1.09	0.28	−0.21 to +0.73
	Range	−2.57 to + 2.09	−1.28 to +1.15			
Nonverbal intelligence (Raven Progressive Matrices; Raven et al., 2003)						
Raw	Mean	36	36	0.24	0.81	−3 to +3
	Range	23 to 49	26 to 55			
Standardized (percentiles)	Mean	64	62			
	Range	17 to 96	20 to 98			
One-minute-real-word reading (Een Minuut test; Brus & Voeten, 1979)						
Raw	Mean	33	62	−8.25	$8.9 \cdot 10^{-12}$	−36 to −21
	Range	5 to 69	31 to 87			
Standardized (norm scores)	Mean	5*	11			
	Range	1* to 11	3* to 15			
Two-minute-nonce-word reading (Klepel, Van den Bos et al., 1994						
Raw	Mean	24	56	−8.97	$5.1 \cdot 10^{-13}$	−39 to −25
	Range	3 to 62	27 to 82			
Standardized (norm scores)	Mean	6*	11			
	Range	1* to 11	7* to 14			
Spelling test (Schoolvaardigheidstoets spelling, Braams & de Vos)						
Raw	Mean	7	20	−12.1	$2.2 \cdot 10^{-16}$	−15 to −11
	Range	0 to 18	13 to 27			
Standardized (percentiles)	Mean	13*	53			
	Range	0* to 59	19 to 94			
Expressive grammatical proficiency (Sentence recall task of the CELF-IV; Semel et al., 2010)						
Raw	Mean	31	59	−9	$5.8 \cdot 10^{-13}$	−34 to −22
	Range	12 to 67	32 to 81			
Standardized (norm scores)	Mean	5*	11			
	Range	1* to 13	3* to 16			

Standardized scores that fall below the normal range are marked with an asterisk* (The normal range included scores from 1 SD below the standardized mean (norm scores: *M* = 10; percentiles: *M* = 50%) to scores 1 SD above the standardized mean, and thus ranged from 8 to 12 (norm scores) or 17% to 86% (percentiles)). The Socioeconomic score is designed such that zero represents the average Dutch socioeconomic status score

females, 28 males) and 35 typically developing children (9 females, 26 males). We refer to Table 1 for a summary of the relevant group characteristics.

Materials

Serial reaction time task

We used a serial reaction time task based on the one used by Lum and Kidd (2012). Children were seated in front of a Microsoft Surface 3 tablet computer screen, with a gamepad controller attached to the computer running E-prime (Version 2.0; 2012) software. A visual stimulus (a cartoon picture of a smiley) appeared repeatedly in one of four marked locations on the screen. These locations were arranged in a diamond shape (Figure 1). We instructed children to press the corresponding button on the gamepad controller as quickly and accurately as possible (see Figure 1). Each stimulus was visible until the child pressed the corresponding button, with a maximum response time of 3 seconds. After the child had responded, there was a 250-millisecond interval before the next stimulus appeared.

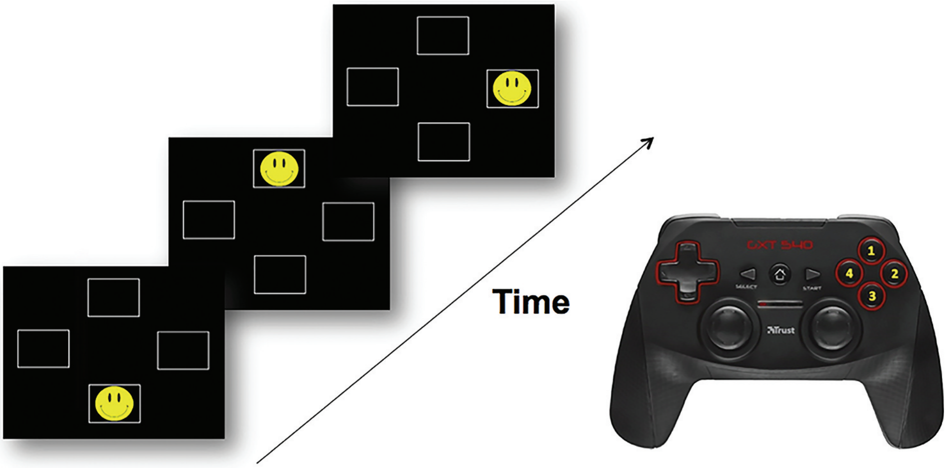


Figure 1. Set up of the serial reaction time task and spatial arrangement of screen locations. Left: a yellow smiley appeared in one of these four marked locations. Right: children were required to press the corresponding buttons on a gamepad controller. The numbers in the gamepad controller buttons correspond to the numbers of the sequence. Number 1 in the sequence corresponds the location at the top of the screen/top button of the gamepad controller. Figure adapted from Van Witteloostuijn (2020), Chapter 4.

Before the start of the real test, we presented children with 28 practice trials to ensure that they understood the task. Unbeknownst to the children, we had divided the stream of stimuli into seven blocks. The first block (20 trials) and sixth block (60 trials) contained trials in a random sequence ("random trials"), whereas the trials in blocks 2 through 5 and in block 7 followed a 10-item deterministic, first-order conditional sequence that was repeated six times (thus 60 trials per block in total). The sequence, where the numbers 1–4 represent the four locations on the screen (Figure 1), was [4, 2, 3, 1, 2, 4, 3, 1, 4, 3]. We refer to these sequenced blocks as "sequence blocks" and to block 6 as the "disruption block". Note that our block structure differs from Lum and Kidd (2012). Their experiment consistent of four sequenced blocks (60 trials each block) and one, final random block.

Sentence recall task

We measured children's productivity of (morpho)syntactic rules with the Sentence Recall Task – a subtest of the Dutch Clinical Evaluation of Language Fundamentals test battery (CELF-4-NL; Semel et al., 2010). In this task, children are instructed to recall sentences with increasing length and complexity. Following the guidelines of the CELF-4-NL, responses are assigned points in relation to the number of errors (e.g., omissions, additions, replacements, substitutions, switches, incorrect markings) made in recalling the sentence. Children receive three points for fully correct recalls, two points for recalls with one error, one point for recalls with two or three errors and zero points for recalls with four or more errors, with a maximum number of 93 points. The task terminates when a child scores zero points on five consecutive recalls. In total, the sentence recall task consists of 31 sentences. The average length of the sentences is 11 words, with the shortest sentence consisting of 5 words and the longest sentence consisting of 19 words. In terms of syllables, the average number of syllables per sentence is 17, with the shortest sentence consisting of 5 syllables and the longest sentence consisting of 30 syllables. The sentences target different Dutch sentence patterns such as subject–verb inversion, as in (1), or subject–verb agreement, as in (2).

(1) Heeft het meisje de bal gevangen?
Has the girl the ball caught?
"Did the girl caught the ball?"
(2) Moeder leest een verhaal voor

```
    Mother reads a story
"Mother reads a story"
```

Furthermore, the sentences differ in whether they target simple or complex Dutch sentence patterns. Following a predefined coding scheme of Van Witteloostuijn (2020) (Chapter 6) 19 out of the 31 sentences used in the sentence recall task can be classified as complex sentences (i.e. 6 passive sentences and 13 subordinate clause sentences).

Procedure

All children took part in our larger study on the relation between statistical learning and grammar and literacy proficiency in children. The total task battery contained more tasks than described here (2 additional statistical learning tasks and a set of additional language tasks and cognitive tasks). The other tasks are described in Lammertink, Boersma, Rispens et al. (2020) and Lammertink, Boersma, Wijnen et al. (2020). All children completed the full task battery, and this took two to four sessions per child. Each child was tested individually. We randomly allocated each child to one of the six different orders in which task administration took place.

Data analysis

Serial reaction time task

We measured accuracy and response time (in milliseconds) of each trial. The accuracy measure served as a sanity check (see Descriptive Results), whereas the response time measure was used to assess children's sensitivity to the underlying structure. We hypothesized that if children were sensitive to the 10-item deterministic sequence, they would show a disruption peak in their response time trajectory, such that their response times in the disruption block (block 6) would be longer than their response times in the preceding and following sequenced blocks (block 5 and block 7). Also, we hypothesized that children with DLD would show a statistical learning deficit, hence that the size of their disruption peak would be smaller than the size of the peak in typically developing children. We obtained an estimate of the size of the disruption peak by selecting children's correct responses to trials in blocks 5, 6 and 7.

In analogy to our earlier work (Lammertink, Boersma, Rispens et al., 2020; Lammertink, Boersma, Wijnen et al., 2020), we normalized children's raw response times so that they could be interpreted as optimally distributed *z* values (see our analysis script at our Open Science Framework (OSF) page: https://osf.io/e9w43/and previous work for normalization procedure). Then, we used a linear mixed effects model that fitted these normalized response times as a function of ternary predictor Block (block 5, block 6, block 7) in interaction with the binary predictor Group (DLD, typically developing children) to assess the size of the statistical learning deficit. The random-effects structure of this model contained by-subject (*N*= 70) and by-position (*N*= 4) random intercepts, by-subject random slopes for the main effect of Block and by-position random slopes for the main effect of Group. The ternary predictor Block was coded such that the first contrast of this predictor ("DisruptionPeak") estimated the size of the disruption peak, with the disruption block coded as +2/3 and with both sequenced blocks coded as -1/3. This predictor disruption peak can be seen as a sanity check, as finding a positive (and statistically significantly different from zero) estimate means that we detected learning, pooled over both groups of children, in our serial reaction time task. The binary predictor Group was coded with DLD as -1/2 and with typically developing children as +1/2. A positive (and statistically significantly different from zero) estimate for the interaction between the first contrast of the predictor Block and the predictor Group was intended to answer our first confirmatory research question, namely whether children with DLD have smaller disruption peaks than typically developing children. We assessed statistical significance of both estimates via 95% Profile confidence intervals and wrote the *get.p.value* function (see Rmarkdown functions script at our OSF) to obtain the corresponding *p* values from the profiles iteratively.

We also computed individual disruption peaks. These individual disruption peaks were used to answer our second confirmatory research question: what is the strength of the correlation between children's performance on the serial reaction time task and their performance on the sentence recall task? We estimate the strength of this correlation for both groups of children separately. In obtaining individual disruption peaks for the children with DLD, we fitted the model described above, but with the predictor Group coded as 0 for DLD and as +1 for typically developing. Then, we extracted with the *ranef* function in R (Bates et al., 2015) participants' (with DLD) random slopes for the predictor disruption peak. We used these random slopes as individual disruption peaks. In obtaining individual disruption peaks for the typically developing children, we undertook the exact same steps, except that the predictor Group was coded +1 for DLD and as 0 for typically developing.

Results experimental study

In what follows, we present only the descriptive results and model estimates that are relevant for our data checks or confirmatory hypothesis testing. All other outcomes are available in the main data analysis script on our OSF project page: https://osf.io/e9w43/. On that page we also made our raw and processed data available.

Descriptive results serial reaction time task

We have no evidence that children with DLD make more (or fewer) errors than typically developing children (pooled over blocks 2 through 7: accuracy children with DLD = 92%; accuracy TD children = 94%, $t= -0.63$, $p= .53$, 95% CI accuracy group difference [−0.054%, +0.028%]). After removing children's incorrect responses and their responses faster than 50 milliseconds (RT < 50 milliseconds: 0.1% in DLD and 0.07% in typically developing children), we calculated the mean raw response times (in milliseconds) with their corresponding standard deviations (in milliseconds) for each block and each group separately (Table 2). These raw response times and standard deviations are computed for ease of exposition only and cannot be used to interpret the strength of effects reported later in this paper or to draw any confirmatory conclusions.

Performance on the serial reaction time task

Though not part of our confirmatory hypothesis testing, we did check whether, pooled over both groups of children, we have evidence that children learned the sequence. The predictor that estimated the size of the learning effect ("DisruptionPeak") was positive and statistically significantly different from zero ($\Delta z = +0.25$, $t= 8.18$, 95% profile CI [+0.19 +0.31, $p= 7.4 \cdot 10^{-9}$]. Children's response times were thus shorter in the fourth training block and recovery block as compared to their response times in the disruption bock. From this we conclude that children can learn the sequence. To obtain an estimate of the maximal standardized effect size for this learning effect, we divided the maximal absolute raw effect size (i.e., the greater absolute bound of the confidence interval) by the residual standard deviation (SD) of the model (residual SD = 0.86). The estimate of the maximal standardized effect size is 0.36 (0.31/0.86). This effect size can be interpreted as a Cohen's d effect size (Cohen, 1988) and as it is > 0.20, but < 0.50 it means that we detected a small to medium sized learning effect.

Table 2. Descriptive mean raw response times and standard deviations (in parentheses), both in milliseconds for the sequenced blocks and disruption blocks for the children with DLD and without DLD separately. DLD = developmental language disorder.

	Block 2 (sequenced)	Block 3 (sequenced)	Block 4 (sequenced)	Block 5 (sequenced)	Block 6 (disruption)	Block 7 (sequenced)
DLD	679 (327)	685 (351)	705 (384)	698 (399)	784 (402)	717 (383)
Typically developing	678 (314)	704 (359)	700 (354)	729 (411)	798 (402)	708 (357)

To answer our first confirmatory research question, we looked at the estimate for the interaction between the predictors DisruptionPeak and Group. This estimate was positive ($\Delta\Delta z = +0.019$): the disruption peak was larger in our typically developing children than in our children with DLD, although the estimate of the group difference is not significantly different from zero (t= 0.32, 95% profile CI [−0.10, +0.14], p=.75). Therefore, we cannot conclude that the size of the disruption peak differs or does not differ between typically developing children and children with DLD (Figure 2). The maximal standardized difference between both groups of children (i.e. the greater absolute bound of the confidence interval for the group-difference divided by the residual standard deviation of the model) is 0.16 (0.14/0.86). In terms of Cohen's d effect size (Cohen, 1988) this effect size is < 0.20, meaning that if a DLD-TD difference exists at all, the difference will be small.

Finally, to further explore whether both groups of children separately showed a statistically significant disruption peak, we fitted two additional models on the exact same data, but with different contrast settings for the predictor Group (with DLD coded as 0 and TD coded as +1 to estimate the disruption peak in DLD, and with DLD coded as +1 and TD coded as 0 to estimate the disruption peak in TD). The estimate for the size of the disruption peaks in both groups of children was positive (DLD: $\Delta z = +0.24$; TD: $\Delta z = +0.25$) and statistically significantly different from zero (DLD: t= 5.56, 95% profile CI [+0.15, +0.32, p= $4.6 \cdot 10^{-7}$, point-estimate effect size: 0.28; Typically developing: t= 6.02, 95% profile CI [+0.17, +0.34, p= $6.2 \cdot 10^{-8}$,point-estimate effect size: 0.29). From this we conclude that both children with DLD and typically developing children learn the sequence and are thus sensitive to the regularities in the input.

Serial reaction time task performance and expressive grammatical proficiency

To answer our second confirmatory research question, we used the *cor.test* function in R (R Core Team, 2018) to compute Pearson correlations between the sizes of children's individual disruption peaks and their scores on the sentence recall task (DLD mean score: 31 points, ranging from 12 to 67 points; Typically developing mean score: 59 points, ranging from 31 to 81 points, Table 1). In both groups, the confidence intervals for the correlation include zero and thus we found no evidence for or

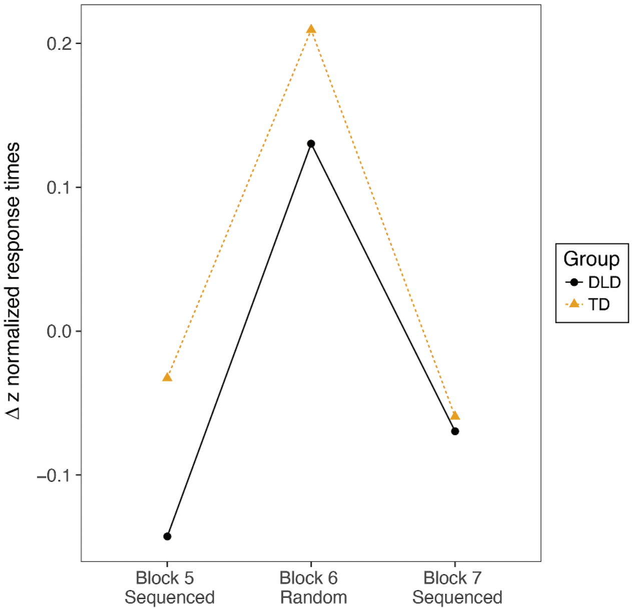


Figure 2. Model estimates of the Normalized response times to the items across block 5 (sequenced), block 6 (random) and block 7 (sequenced). Normalized response times are plotted for the children with DLD (black line) and typically developing children (orange line) separately.

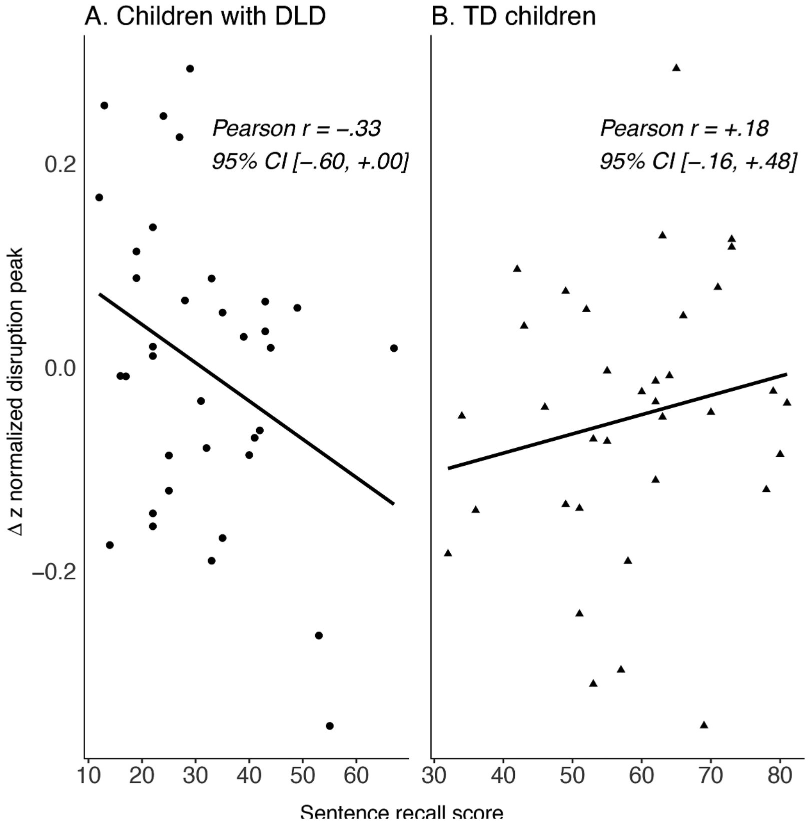


Figure 3. Descriptive visualization of the correlation between the size of children's individual disruption peak (centered and scaled, vertical axis) and their average points obtained on the sentence recall task from the CELF (centered and scaled, horizontal axis). The correlation for children with DLD is plotted with black circles and on the left side. The correlation for typically developing children in plotted with black triangles and on the right sight. Each circle and triangle represent the correlation for an individual child. TD = typically developing.

against a relationship between children's size of the disruption peak and their score on the sentence recall task (DLD: *r* (33) = −.33, 95% CI [−.60, +.00]; TD: *r* (33) = +.18, 95% CI [−.16, +.48], Figure 3).

Split-half reliability disruption peak measure

We also assessed internal consistency of the disruption peak measure as an individual of measure serial reaction time task performance. We computed the split-half reliability (Spearman-Brown corrected Pearson correlation between the size of children's individual disruption peak for even items and the size of children's individual disruption peak for odd items). The split-half reliability is 0.62, with the 95% confidence interval ranging from 0.38 to 0.76. Ideally, the split-half reliability should reach the psychometric standard of *r*= .80 (e.g., Nunnally & Bernstein, 1994; Streiner, 2003).

Discussion experimental study

The experiment was designed to assess the strength of the association between serial reaction time performance and expressive grammar in children with and without DLD. Additionally, we aimed to replicate previous findings showing that children with DLD are less sensitive to structural regularities in the visuomotor domain as compared to their typically developing peers (see meta-analysis Lum et al., 2014). Therefore, we used a serial reaction time task design that is commonly used to assess the difference in performance between children with and without DLD. The task that we used was based on the one used by Lum and Kidd (2012). Lum and Kidd did not compare serial reaction time task performance between children with and without DLD, but task designs similar to the set-up of their task (see Introduction) have been used to assess the presence of a visuomotoric statistical learning deficit in children with DLD (Clark & Lum, 2017; Conti-Ramsden et al., 2015; Hsu & Bishop, 2014; Park et al., 2018). Unexpectedly, we observed that both groups of children were sensitive to the structural regularities, and we found no evidence for or against a difference in sensitivity between children with and without DLD. Given the similarities in design and number of children tested, we have no clear explanation for the small DLD–TD difference observed in our study as compared to the outcome of the meta-analysis by Lum et al. (2014).

In a second step we investigated the strength and existence of an association between performance on the serial reaction time task and children's sentence recall performance. We found no evidence for or against an association in children with and without DLD. While we have no reason to believe that the power of our study was too low to detect differences in learning at the *group* level (we detected a small to medium sized learning effect in both groups of children), it was a priori uncertain if the number of participants was sufficient to detect an association between children's serial reaction time task performance and their sentence recall performance. We now observe that the confidence intervals of our associations ranged from −.60 to +.00 in children with DLD and from −.16 to +.48 in typically developing children. The estimated upper bounds of the "standardized" effect sizes for these associations are $R^2 = .36$ ($.60^2$) and $R^2 = .23$ ($.48^2$). This means that the true effects lie between 0 and medium (as the standardized effect sizes are >0.20 but <0.50, Cohen, 1988). Only if these ranges had been small, we could have concluded that we tested sufficiently children to detect the associations.

Another potential explanation for our null result on the association between serial reaction time performance and grammatical proficiency may concern the reliability of the disruption peak as an *individual* measure of statistical learning in the serial reaction task (see Arnon, 2019; Siegelman, Bogaerts, Frost et al., 2017; West et al., 2017 for discussions of the low psychometric properties of the currently available measures to assess individual differences in statistical learning). Our disruption peak measure had a split-half reliability of 0.62, with a 95% confidence interval ranging from 0.38 to 0.76. This is a relatively wide confidence interval suggesting, indeed, that the internal consistency (one of the measures of reliability) of our individual disruption peak measure may not be as high as preferred and may not approach the psychometric standard of $r = .80$ (Nunnally & Bernstein, 1994; Streiner, 2003).

To put our null result on the association between serial reaction time performance and grammatical proficiency within the context of previous work on this topic, we decided to conduct a meta-analysis. This meta-analysis also allows us to assess the existence of a potential publication bias. The meta-analysis is discussed in the following sections.

Study 2: Meta-analysis

Methods meta-analysis

We used the Preferred Reporting Items for Systematic Reviews and Meta-analysis statement to organize the current meta-analysis (Moher et al., 2009). Effect size calculations and statistical analyses on the effect size measures were done in R (R Core Team, 2018).

Literature search

A first systematic search was conducted by the first author of this paper in February 2018. The search was conducted in five different sources: PubMed, PsycINFO, Education Resources Information Center (ERIC), Linguistics and Language Behavior abstracts (LLBA) and Open Access Theses and Dissertations (OATD). In addition, the first author also contacted experts in the field (via the

LINGUIST List and via the Cogdevsoc list) with requests for access to unpublished data. Altogether, this first search yielded 93 unique articles (91 hits via the databases and 2 hits via the mailing lists; see flow chart in Appendix A). A second search in PubMed, PsycINFO and OATD, which served as a reliability check, was done by a research assistant in September 2018. This second search yielded 13 additional potentially relevant unique articles that were not in the output of the first search. Finally, a third search was conducted by another research assistant in January 2019. This third search was conducted as we realized that our initial query focused on studies that included people with DLD/specific language impairment and that therefore, we might have missed articles on serial reaction time task performance in typically developing children. This third search yielded 11 additional potentially relevant unique articles. Thus, in total we screened 115 unique articles on their title and abstract. If, by screening the title and/or abstract, it became clear that the study did not meet the inclusion criteria for the meta-analysis (see Inclusion Criteria and Study Selection), then the study was excluded. For 49 articles or datasets, we read the methods and result sections carefully in order to decide whether or not the study met the inclusion criteria. Eventually, 18 articles (15 published articles, 1 preprint and 2 dissertations) met our inclusion criteria and were included in our database (see Sample Description). See our OSF page for Excel spreadsheets with information on why studies were eventually included or excluded for analysis.

Inclusion criteria and study selection

Studies were eligible and included in our meta-analysis if they met all of the following criteria:

(1) The study involved the use of a serial reaction time task in the visuomotor domain, comprising nonlinguistic stimuli.
(2) The study reported on a measure of children's grammatical proficiency, or it became clear that the authors had information on children's grammatical proficiency.
(3) The study involved typically developing children and/or children with DLD (or specific language impairment) between four and twelve years old. Please note that for studies in which typically developing children were compared to a clinical population other than DLD (e.g., children with dyslexia, autism spectrum disorder, deaf children), we included only the results from the typically developing children. As the criteria for having DLD varied between studies, we decided that in order to be classified as DLD, the following criteria would have to be met: (a) children were identified as having DLD using scores on a (standardized) language test battery that differentiated between children with and without language impairment, that (b) the children with DLD and their typically developing peers were matched on nonverbal intelligence, and that (c) children had no history of a neurological disorder and/or emotional delay.
(4) Finally, for the present paper, we only included studies that were conducted before September 2018. Our database is community-augmented, however, meaning that it is accessible online via our OSF project page and open to updates (Tsuji et al., 2014).

Sample description

The final sample includes 56 effect sizes pertaining to correlations between an index of serial reaction time performance and grammatical proficiency. Twenty-nine of these 56 effect sizes are correlations with an expressive grammar index. The other 27 effect sizes are correlations with a receptive grammar index. We could not include both grammar types in one meta-analysis, because in the majority of studies the receptive grammatical proficiency and expressive grammatical proficiency measures came from the same individual children (violating the assumption of independence). A solution to this problem could have been to calculate a synthesized effect size of both measures. We had two reasons not to do so, however. First, synthesizing these effect sizes requires knowledge of the correlation between children's expressive grammatical proficiency and their receptive grammatical proficiency.

These correlations are often not reported. Second, it is yet unknown whether the strength of the association differs between expressive and receptive measures of grammatical proficiency. Because our experimental study assesses the association with expressive grammatical proficiency, we decided to focus on this type of grammatical proficiency. Note, however, that we explored the correlation with receptive grammatical proficiency in a separate meta-analysis. Details of this exploratory analysis can be found in the Rmarkdown file on our OSF-page.

At this point it should be noted that also within the set of expressive grammar correlations, some studies reported more than one correlation from the same sample of children (e.g., they reported both a correlation with sentence recall and with sentence formulating). In most of these cases, we did obtain the correlation between the two measures of expressive grammar via personal communication with the authors and thus, we could synthesize these effect sizes (see Effect Size Computation and Synthesized Effect Sizes). After selecting and/or synthesizing effect sizes that came from the same sample of children (see Effect Size Computation and Synthesized Effect Sizes), the final dataset contained 19 unique correlations between expressive grammar (indexed by a sentence recall or sentence completion task) and serial reaction time task performance (first-order conditional sequence: *N*= 13; second-order conditional: *N*= 6) in children with DLD (*N*= 8 effect sizes, 139 children with DLD) and in typically developing children (*N*= 11 effect sizes, 573 typically developing children).

Effect size computation

From each study, we extracted the relevant correlation coefficients. If needed, we synthesized effect sizes that came from the same sample of children (see Synthesized Effect Sizes). The extracted correlations were transformed into Fisher *z* values with their corresponding variances (Borenstein et al., 2009, p. 42, Formula A, Formula B in appendix B). All studies, except the study of Hani (2015), reported Pearson *r* correlations. The study by Hani (2015) reported a Kendall's *tau* correlation and therefore we first transformed this correlation into Pearson *r* (Formula C, Appendix B) before transforming it into Fisher *z*.

Synthesized effect sizes

There were nine articles that reported multiple correlations between serial reaction time task performance and expressive grammatical proficiency in the same group of children. These multiple correlations were reported either because children performed multiple serial reaction time tasks at different timepoints, or because the authors obtained multiple measures of children's expressive grammatical proficiency (e.g., children did both a sentence recall and a sentence formulation task). As already mention above, we cannot include correlations that come from the same group of children in one meta-analysis, as that would violate the assumption of independence. Therefore, we either selected (a) only one of the correlations reported or (b) we computed (or obtained) a synthesized effect size across the multiple correlations reported. We chose option (a) if the multiple correlations were reported for different timepoints, and option (b) if multiple measures of expressive grammar were reported. In the case of solution (a), we decided to select only the correlation reported for the child's first serial reaction time task session (Desmottes et al., 2016a; Gabriel et al., 2012; West et al., 2020, 2017). This also meant that we eventually had to exclude Desmottes, Meulemans et al. (2017), because they only reported a SRT-Grammar correlation for the children's third SRT session. In the case of solution (b), we computed (or obtained) a synthesized (combined) effect size, and its associated synthesized variance (formulas D and E; Borenstein et al., 2009, p. 227; Desmottes et al., 2016a; Desmottes, Maillart et al., 2017; Kidd, 2012; Kidd & Kirjavainen, 2011; Obeid, 2017; Park et al., 2018). The resulting synthesized effect sizes are reported in Table 3. Kidd and Kirjavainen (2011) report on correlations between children's serial reaction time task

Table 3. Overview of studies for which we computed a synthesized effect size (i.e. average combined correlation between children's performance on the serial reaction time task and their scores on the two measures of grammar proficiency used). Computation of these synthesized effect sizes requires knowledge of the correlation between the two measures of grammar proficiency. These correlations are reported in the first column. *Correlation between lax test and lax post test (Kidd, 2012) **Correlation between children's SRT rebound score (raw scores) and their overall accuracy score on the past tense task (raw accuracy scores for all six categories combined). Obtained via personal communication with Evan Kidd.

	Correlation (Pearson *r*) between grammar index 1 and grammar index 2		Synthesized effect size (Pearson *r*) with its corresponding variance (in parentheses)	
	DLD	TD	DLD	TD
Kidd (2012)*	n/a	+.10	n/a	+.21 (+.0050)
Kidd and Kirjavainen (2011)**	n/a	n/a	n/a	+.080 (+.0083)
Park et al. (2018)	+.67	+.33	−.24 (+.042)	.00 (+.019)
Obeid (2017)	n/a	+.57	n/a	−.043 (+.013)
Desmottes, Meulemans, & Maillart, 2016a	+.82	+.38	−.28 (+0.039)	−.073 (+.023)
Desmottes, Maillart et al., 2017	+.82	+.38	+.24 (+0.047)	+.051 (+.043)

performance and children's accuracy on six different Finnish past tense categories. For the purpose of this meta-analysis, Kidd provided us (personal communication) the correlation between children's serial reaction time task performance (SRT rebound; raw scores) and children's average raw accuracy score pooled over the six past tense categories. For the remaining studies, we calculated the synthesized effect sizes ourselves. The calculation of these synthesized effect sizes required knowledge of the correlation between the different measures of expressive grammar. For Kidd (2012), Park et al. (2018) and for Obeid (2017), we obtained these correlations from the authors (see our OSF page). Unfortunately, we did not obtain this information for the Desmottes et al. (2016a) and Desmottes, Maillart et al. (2017) papers. Therefore, we took these correlations from another paper by the same authors (Desmottes et al., 2016b) in which they did report the correlations, although for different samples of children.

Data analysis and coding of moderator variables

The main aim of the meta-analysis was to assess the strength of the relationship between serial reaction time task performance and expressive grammatical proficiency in primary-school aged children. We set out to answer this confirmatory research question with a hierarchical meta-analytic random effects model (see *rma.mv* function from the *metafor* package, Version 2.0.0 in R, Viechtbauer, 2010) in which we fitted the mean weighted correlation as a function of the binary moderator Group, with DLD coded as -1/2. and with typically developing coded as +1/2. The random-effects structure contained a random intercept for Paper (*N*= 13). Simultaneously we also explored whether the mean weighted correlation is stronger in typically developing children than in children with DLD (as the procedural learning deficit hypothesis may predict; Ullman & Pierpont, 2005). A positive (and statistically different from zero) estimate for the predictor Group may be a preliminary indication that this hypothesis is true.

In a secondary step, we explored whether the mean weighted correlation (when controlling for group status) varied by sequence type (first-order conditional versus second-order conditional) or Age. These exploratory analyses were conducted through model comparisons. Generally, if moderators affect the strength of the correlation, adding them to the model will result in better model fits. With the first model comparison, we compared the "Group-model" (as specified above) to the "Group-Sequence" model. In this Group-Sequence model, the effect size is fit as a function of the binary moderator Group, the binary moderator Sequence Type (with first-order conditional coded as +1/2 and with second-order conditional coded as -1/2) and the interaction between both moderators. The second model comparison compared the Group model to the "Group-Age" model in which the effect size is fitted as a function of the binary moderator Group, the continuous predictor Age in months (centered and scaled, ranging from −2.03 to +0.90) and the interaction between Group and Age.

Results of the meta-analysis

Publication bias

To assess the presence of a publication bias in the present meta-analysis, we analyzed funnel plot asymmetry (Egger et al., 1997) with a linear regression on our funnel plot (Figure 4). Visual inspection of our funnel plot suggests that the effect sizes are symmetrically distributed and therefore publication bias seems unlikely. Using the *regtest* function in the *metafor* package (Version 2.0.0) of the statistical programming language R (Viechtbauer, 2010), we found no evidence for or against funnel plot asymmetry (publication bias) in our sample (z= −0.94, p= .35).

Confirmatory meta-analysis

The model outcome provided no evidence for or against a correlation between serial reaction time task performance and expressive grammatical proficiency in the pooled group of children (Fisher z = 0.072, SE = 0.051, z= 1.41, p= .16, 95% CI [−0.028, +0.17]). For ease of interpretation, we also transformed the Fisher z estimate and its 95% confidence interval back into Pearson r values: the corresponding values are: r = .072, 95% CI [−.028, +.17].

Exploratory analyses

In addition to assessing the strength of the correlation between serial reaction time performance and expressive grammatical proficiency, we also explored whether the strength of this relationship differed between children with and without DLD. The model outcome of the predictor estimated that the strength of the relationship is stronger in typically developing children than in children with DLD. However, the estimate of the group difference was not significantly different from zero (Fisher z = +0.079, SE = 0.10, z= 0.77, p= .44, 95% CI [−0.12, +0.28]; Pearson r = +.079, 95% CI [−.012, +.27])

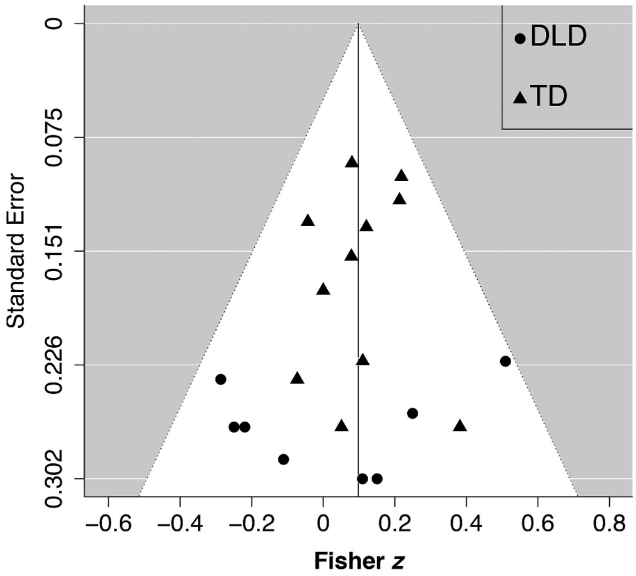
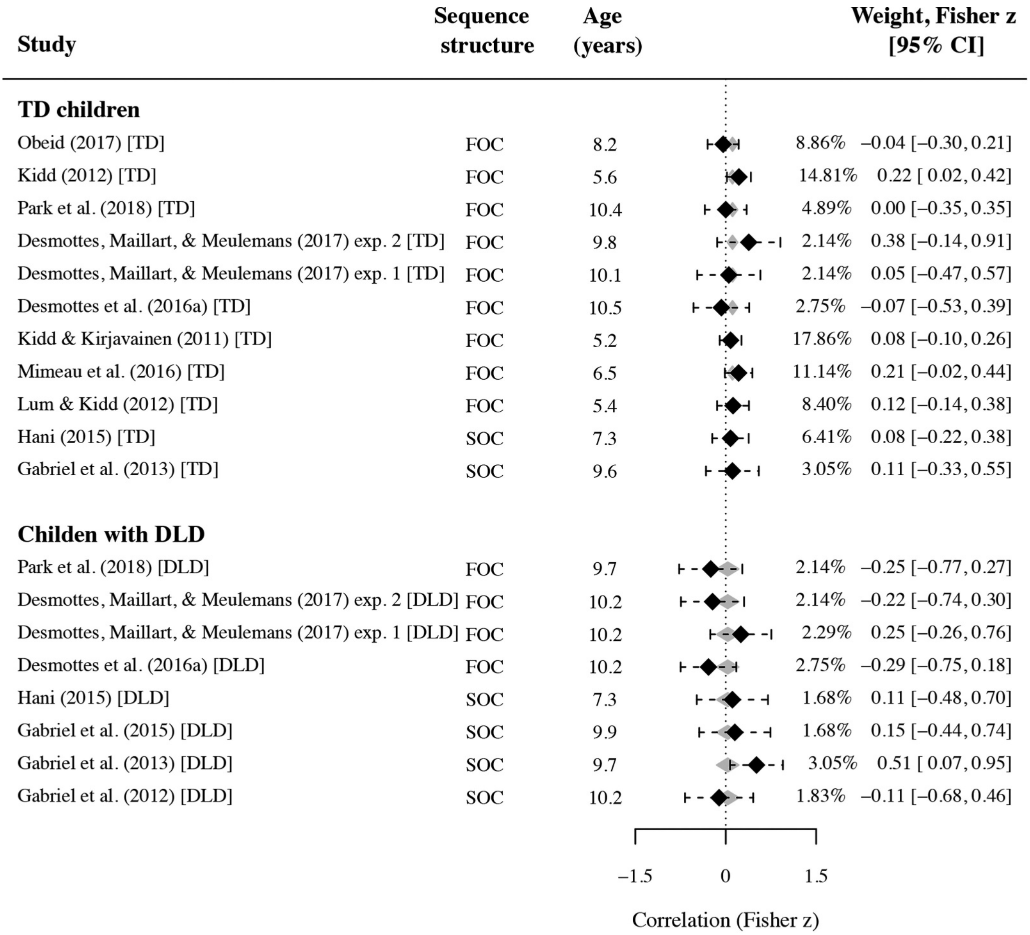


Figure 4. Funnel plot showing standard error of the effect size Fisher z as a function of the effect size. The vertical line indicates the mean weighted correlation. Dots in black are individual effect sizes from children with DLD, dots in yellow represent individual effect sizes from typically developing children. The triangle-shaped unshaded region represents a pseudo confidence interval region with bounds equal to ± 1.96 SE.



Study	Sequence structure	Age (years)	Weight, Fisher z [95% CI]
TD children			
Obeid (2017) [TD]	FOC	8.2	8.86% −0.04 [−0.30, 0.21]
Kidd (2012) [TD]	FOC	5.6	14.81% 0.22 [0.02, 0.42]
Park et al. (2018) [TD]	FOC	10.4	4.89% 0.00 [−0.35, 0.35]
Desmottes, Maillart, & Meulemans (2017) exp. 2 [TD]	FOC	9.8	2.14% 0.38 [−0.14, 0.91]
Desmottes, Maillart, & Meulemans (2017) exp. 1 [TD]	FOC	10.1	2.14% 0.05 [−0.47, 0.57]
Desmottes et al. (2016a) [TD]	FOC	10.5	2.75% −0.07 [−0.53, 0.39]
Kidd & Kirjavainen (2011) [TD]	FOC	5.2	17.86% 0.08 [−0.10, 0.26]
Mimeau et al. (2016) [TD]	FOC	6.5	11.14% 0.21 [−0.02, 0.44]
Lum & Kidd (2012) [TD]	FOC	5.4	8.40% 0.12 [−0.14, 0.38]
Hani (2015) [TD]	SOC	7.3	6.41% 0.08 [−0.22, 0.38]
Gabriel et al. (2013) [TD]	SOC	9.6	3.05% 0.11 [−0.33, 0.55]
Childen with DLD			
Park et al. (2018) [DLD]	FOC	9.7	2.14% −0.25 [−0.77, 0.27]
Desmottes, Maillart, & Meulemans (2017) exp. 2 [DLD]	FOC	10.2	2.14% −0.22 [−0.74, 0.30]
Desmottes, Maillart, & Meulemans (2017) exp. 1 [DLD]	FOC	10.2	2.29% 0.25 [−0.26, 0.76]
Desmottes et al. (2016a) [DLD]	FOC	10.2	2.75% −0.29 [−0.75, 0.18]
Hani (2015) [DLD]	SOC	7.3	1.68% 0.11 [−0.48, 0.70]
Gabriel et al. (2015) [DLD]	SOC	9.9	1.68% 0.15 [−0.44, 0.74]
Gabriel et al. (2013) [DLD]	SOC	9.7	3.05% 0.51 [0.07, 0.95]
Gabriel et al. (2012) [DLD]	SOC	10.2	1.83% −0.11 [−0.68, 0.46]

Figure 5. Forest plot showing overall and individual mean weighed effect sizes (Fisher z), divided per group. The shaded diamonds represent the mean weighed effect size per group (DLD or typically developing). FOC = first order conditional; SOC = second order conditional.

We also explored whether the mean weighted correlation between serial reaction time performance and expressive grammar, controlled for Group status (DLD versus typically developing) differed as a function of sequence type (first-order conditional versus second-order conditional) or age. Model comparisons revealed that we cannot conclude that this is the case. Neither the Group model versus Group–Sequence model comparison (*p*= .17) nor the Group model versus Group–Age model comparison (*p*= .54) was significantly different from zero.

Finally, as explained in our Sample description section, we also explored the correlation between receptive grammar and serial reaction time task performance in a separate meta-analysis. Details of this exploratory analysis can be found in the Markdown file on our OSF-page. There was no evidence for (or against) the presence of a publication bias in the set of receptive grammar studies (*z*= .15, *p*= .88). Also, the model outcome provided no evidence for or against a correlation between serial reaction time task performance and receptive grammatical proficiency in the pooled group of children (Fisher *z* = 0.051, *SE* = 0.041, *z*= 1.24, *p*= .22, 95% CI [−0.030, +0.13]), nor that the strength of the correlation differs (or does not differ) between children with and without DLD (Fisher *z* = +0.013, *SE* = 0.080, *z*= 0.17, *p*= .87, 95% CI [−0.14, +0.17]).

Discussion of the meta-analysis

The present meta-analysis provided a quantitative overview of published and unpublished studies on the association between serial reaction time performance and expressive grammatical proficiency in children with and without DLD. Summarizing over 19 unique correlations that collectively examined 139 children with DLD (8 effect sizes) and 573 typically developing children (11 effect sizes), we found no evidence for or against the existence of an association between serial reaction time task performance and expressive grammatical proficiency in children with and without DLD. The declarative memory compensation hypothesis (Ullman & Pullman, 2015) claims that the correlation is smaller in children with DLD as compared to their typically developing peers, which may result in a weaker overall correlation in the pooled group of children. Therefore, we also assessed whether the mean weighted correlation is smaller in children with DLD than in typically developing children. The latter could not be concluded. In the General Discussion we discuss some factors that may have contributed to these inconclusive results.

In the second part of our meta-analysis we further explored whether the strength of the proposed association differed as a function of sequence structure (first-order conditional versus second-order conditional) or age. We found no evidence, however, that these factors did or did not moderate the strength of the association.

General discussion

The main aim of the present study was to provide an in-depth overview and an evaluation of the relation between serial reaction time performance and expressive grammatical proficiency in children with and without DLD. In doing so, we first presented the results of our experimental study, which was a conceptual replication of previous work on the presence of a visuomotoric statistical learning deficit in children with DLD. Unexpectedly, we cannot conclude that we replicated (or did not replicate) previous work on this topic. We find no evidence for (or against) the existence of visuomotoric statistical learning deficit in children with DLD. We observed that both children with DLD and typically developing children learned the sequential structure, suggesting that children with DLD, like typically developing children, are sensitive to sequential regularities in the visuomotor domain. Also, when using the size of the disruption peak as an individual measure of visuomotoric statistical learning, we found no evidence for or against an association between statistical learning and expressive grammatical proficiency in our sample of children with and without DLD.

In an attempt to explain these null results, we realized that there was no clear consensus on (a) the existence and strength of the proposed association and (b) to what extent the relation is weaker in children with DLD than in typically developing children (as proposed, for example, by the declarative memory compensation hypothesis Ullman & Pullman, 2015). This motivated us to conduct the meta-analysis described in the second part of the paper. The outcomes provide no evidence for (or against) an association between serial reaction time performance and expressive grammar, nor evidence that the strength of this association differs between children with DLD and without DLD. Also, our meta-analysis provided no evidence for the existence of a publication bias, and therefore we cannot conclude that the outcomes of the meta-analysis are influenced (or not) by publication bias (note that a publication bias has been observed in the literature on statistical learning in children with dyslexia by Schmalz et al., 2017 and by; Van Witteloostuijn et al., 2017).

There are various factors that may have contributed to these inconclusive results. Firstly, and as mentioned in the Discussion of our experimental study, they may be partially the result of psychometric shortcomings in the currently available measures to assess individual differences in statistical learning (Arnon, 2019; Siegelman, Bogaerts, Frost et al., 2017; West et al., 2017). It goes beyond the scope of this paper, to again discuss these issues extensively but it should be reiterated that there is a methodological debate within the statistical learning literature on how to reliably measure individual differences in statistical learning. This debate concerns the improvement of the psychometric properties of individual measures of statistical learning (for suggestions on novel measures of statistical learning see, for example: Isbilen et al., 2017; Kidd et al., under review; Lammertink et al., 2019; Siegelman, Bogaerts, Frost et al., 2017), the use of samples that are too small to detect differences in learning between children with and without DLD or to detect associations between statistical learning and measures of linguistic proficiency (West et al., 2017), as well as more theoretical discussion on what additional factors impact statistical learning performance and the detection of an association between statistical learning and other cognitive abilities. Questions explored related to this latter point are to what extent statistical learning can be dissociated from other cognitive processes as attention and working memory (Arciuli, 2017; West et al., 2020), but also whether statistical learning and the reliability of the different measurements of statistical learning are age-dependent (for reviews on this topic see Arciuli, 2017; Krogh et al., 2013; Zwart et al., 2019).

Secondly, studies on the relation between statistical learning and other cognitive processes often spend very little time discussing the theoretical motivation behind the selection of their tasks (as commented on by Siegelman, Bogaerts, Christiansen et al., 2017). As a consequence, the sequential structure targeted in the statistical learning tasks is often only tangentially related to structure relevant for the linguistic ability that researchers try to predict with their task, let alone to how children acquire language in real life. At the same time, the tasks of grammatical proficiency also differ in the grammatical structures that they employ, with most standardized measures (like the sentence recall task) assessing children's knowledge of a mixture of grammatical structures. The sentence recall task (Semel et al., 2010), for example, measures children's knowledge of different sentence types (e.g., passives, declaratives, relative clause constructions), different morphosyntactic processes (subject–verb agreement, past-tense production, pluralization) and likely also other cognitive processes such as working memory (Frizelle et al., 2017). The pattern that needs to be learned in the serial reaction time task may not be relevant in predicting sensitivity to all these different sentence types and morphosyntactic constructions (Kidd & Arciuli, 2016; Mimeau et al., 2016; Misyak & Christiansen, 2012; Wilson et al., 2018). Also, among the studies included in our meta-analysis we observe large variability in the measures of grammatical proficiency used. It is debatable whether all these different measures tap into a similar underlying construct and whether differences in task demands modulate the strength of the correlation with serial reaction time task performance. For future studies it may thus be interesting to focus on more specific grammatical constructs, preferable grammatical structures that, similarly to the serial reaction time task, require sensitivity to sequential information (e.g., children's sensitivity to and comprehension of (non)canonical word order; Montgomery et al., 2017).

Finally, there may be a discrepancy in how acquired knowledge is measured in statistical learning tasks versus how acquired knowledge is measured in grammatical proficiency tasks. In the present sample of studies, the measures used to assess visuomotoric statistical learning are all processing-based (i.e. based on response times), whereas the measures used to assess grammatical proficiency are all, except for Clark and Lum (2017), accuracy-based. Processing-based measures may be more sensitive to implicit knowledge representations, whereas accuracy-based measures may be more sensitive to explicit knowledge representations (Franco et al., 2015; Isbilen et al., 2017; Misyak et al., 2010). This discrepancy may complicate the detection of an association between the two cognitive systems. In this context, it may be particularly interesting to follow publications on a newly developed measure of auditory statistical learning by Isbilen et al. (2017) and Kidd et al. (under review). Using a serial recall as a measure of auditory statistical learning, these researchers show that adults (Isbilen et al.) and children (Kidd et al.) repeat trained sequences of syllables better than foils. Given the similarity in how acquired knowledge is measured with this serial recall measure of statistical learning and how language proficiency is measured with the sentence recall task (e.g., both measures are based on recall), it would be interesting to investigate whether stronger correlations with sentence recall performance are observed for the "recall" measures of statistical learning as compared to the currently available measures of statistical learning.

Conclusion

Neither our own experiment nor our meta-analysis provides any evidence for the existence of an association between serial reaction time performance and expressive grammatical proficiency in children with and without DLD. The confidence interval of the meta-analysis (Pearson's *r* from −.028, to + .17) is compatible with a non-existent association, but also with a small-to-medium-sized association. We speculate that such an association may exist only if (a) the targeted structure in the statistical learning task is meaningfully related to the target structure in the grammatical proficiency task *and* (b) both measures represent the same represent the same response type of the participant. Overall, it may be that visuomotoric statistical learning is associated with expressive grammar but that we encountered methodological problems in its detection. Taken together, we cannot claim yet that a visuomotoric statistical learning deficit is or is not associated with the language problems observed in children with DLD.

Acknowledgments

This work is part of the research programme Vernieuwingsimpuls – Vidi (examining the contribution of procedural learning to grammar and literacy in children) with the project number 27689005, which is (partly) financed by the Dutch Research Council (NWO). For the experimental part of this paper, we extend our gratitude to all children who participated and to their parents, teachers, and speech therapists who facilitated the children's participation. More specifically, we would like to thank Viertaal special education in Almere, Amsterdam, Purmerend, and Schagen; The Royal Dutch Auris Group in Breda, Haarlem, Leiden, and Tilburg; The Royal Dutch Kentalis in Nijmegen; Pento in Amersfoort, and stichting Hoormij for their help in the recruitment of children with DLD. Also, we would like to thank the four primary schools (Binnenmeer, 'T Blokhuus, Startnest, and Wheermolen) that participated. We also thank Evan Kidd and Jarrad Lum for sharing their serial reaction time task and Dirk Jan Vet for the technical implementation of the experiment. Finally, we thank Iris Broedelet, Sascha Couvee, Darlene Keydeniers and Maartje Hoekstra (test assistants). For the meta-analysis we thank Hailin Hao and Ellen Collée (literature search and coding of variables). Also, we thank Dorothy Bishop, Lise Desmottes, Sengottuval Kuppuraj, Chirstelle Maillart, Catherine Mimeau, Rita Obeid, Jisook Park, Sybren Spit and Gillian West for providing additional data and information that enabled us to conduct the meta-analysis.

Disclosure statement

The authors have no (financial) conflict of interest in the subject matter or materials discussed in the manuscript.

Funding

This work was supported by the Dutch Research Council (NWO) [27689005].

ORCID

Imme Lammertink http://orcid.org/0000-0001-6625-4108

References

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160058. https://doi.org/10.1098/rstb.2016.0058

Arciuli, J. (2018). Reading as statistical learning. *Language, Speech, and Hearing Services in Schools*, *49*(3S), 634–643. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0135

Arnon, I. (2019). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modalities. *Behavioral Research Methods, 52,* 68–81. https://doi.org/10.3758/s13428-019-01205-5

Bates, D., Maechler, M., Bolker., B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i0

Bishop, D., Snowling, M., Thompson, P., & Greenhalgh, T. (2017). Phase 2 of CATALISE: A multinational multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, *58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual meeting of the cognitive science society* (pp. 124–129). Austin, TX: Cognitive Science Society.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons, Ltd.

Braams, T., & de Vos, T. (2015). *Schoolvaardigheidstoets spelling [dutch spelling test; measurement instrument]*. Boom test uitgevers.

Brus, B., & Voeten, M. (1979). *Een-minuut-test [one minute test; measurement instrument]*. Pearson.

Clark, G., Barham, M., Ware, A., Plumridge, J., O'Sullivan, B., Lyons, K., & Lum, J. (2019). Continuous theta-burst stimulation reveals dissociable sequence learning networks. *Behavioural Neuroscience*, *133*(4), 341–349. https://doi.org/10.1037/bne0000299

Clark, G., & Lum, J. (2017). Procedural memory and speed of grammatical processing: Comparison between typically developing children and language impaired children. *Research in Developmental Disabilities*, *71*, 237–247. https://doi.org/10.1016/j.ridd.2017.10.015

Cohen, A., Ivry, R., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 17–30. https://doi.org/10.1037/0278-7393.16.1.17

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associations.

Cohen, J. (1992). A power primer. *Psycholinguistic Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry*, *42*(6), 741–748. https://doi.org/org/10.1111/1469-7610.00770

Conti-Ramsden, G., Ullman, M., & Lum, J. (2015). The relation between receptive grammar and procedural, declarative, and working memory in specific language impairment. *Frontiers in Psychology*, *6*(1090), 1–11. https://doi.org/10.3389/fpsyg.2015.01090

Desmottes, L., Maillart, C., & Meulemans, T. (2017). Memory consolidation in children with specific language impairment: Delayed gains and susceptibility to interference in implicit sequence learning. *Journal of Clinical and Experimental Neuropsychology*, *39*(3), 265–285. https://doi.org/10.1080/13803395.2016.1223279

Desmottes, L., Meulemans, T., & Maillart, C. (2016b). Implicit spoken words and motor sequences learning are impaired in children with specific language impairment. *Journal of the International Neuropsychological Society, 22*, 520–529. https://doi.org/10.1017/S135561771600028X

*Desmottes, L., Meulemans, T., & Maillart, C. (2016a). Later learning stages in procedural memory are impaired in children with specific language impairment. *Research in Developmental Disabilities*, *48*, 53–68. https://doi.org/10.1016/j.ridd.2015.10.010

*Desmottes, L., Meulemans, T., Patinec, M.-A., & Maillart, C. (2017). Distributed training enhances implicit sequence acquisition in children with specific language impairments. *Journal of Speech, Language and Hearing Research*, *60*(9), 2636–2647. https://doi.org/10.1044/2017_JSLHR-L-16-0146

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

E-prime (version 2.0) [computer software]. (2012). Psychology Software Tools.

Erickson, L., & Thiessen, E. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108. https://doi.org/10.1016/j.dr.2015.05.002

Evans, J., Saffran, J., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189)

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation: A new measure of statistical learning in speech segmentation. *Experimental Psychology*, *62*(5), 346–351. https://doi.org/10.1027/1618-3169/a000295

Frizelle, P., O'Neill, C., & Bishop, D. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language*, *44*(6), 1435–1457. https://doi.org/10.1017/S0305000916000635

*Gabriel, A., Stefaniek, N., Maillart, C., Schmitz, X., & Meulemans, T. (2012). Procedural visual learning in children with specific language impairment. *American Journal of Speech-Language Pathology*, *21*(4), 329–341. https://doi.org/10.1044/1058-0360(2012/11-0044)

*Gabriel, A., Maillart, C., Stefaniek, N., Lejeune, C., Desmottes, L., & Meulemans, T. (2013). Procedural learning in specific language impairment: Effects of sequence complexity. *Journal of the International Neuropsychological Society*, *19*(3), 164–271. https://doi.org/10.1017/S1355617712001270

*Gabriel, A., Meulemans, T., Parisse, C., & Maillart, C. (2015). Procedural learning across modalities in French-speaking children with specific language impairment. *Applied Psycholinguistics*, *36*(3), 747–769. https://doi.org/10.1017/S0142716413000490

Hamrick, P., Lum, J., & Ullman, M. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the national academy of sciences of the Unites States of America*, 115, 1487–1492. https://doi.org/10.1073/pnas.1713975115

*Hani, H. (2015). *Language-impaired children with Autism spectrum disorders and children with specific language impairment: Similar language abilities but distinct memory profiles* (Doctoral dissertation submitted to the School of Communication Sciences and Disorders). McGill University.

Hsu, H., & Bishop, D. (2014). Sequence-specific procedural learning deficits in children with specific language impairment. *Developmental Science*, *17*(3), 352–365. https://doi.org/10.1111/desc.12125

Isbilen, E., McCauley, S., Kidd, E., & Christiansen, M. (2017, July). *Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. Paper presented at the 39th Annual meeting of the cognitive science society*, London, UK.

*Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*(1), 171–184. https://doi.org/10.1037/a0025405

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184–193. https://doi.org/10.1111/cdev.12461

Kidd, E., Arciuli, J., Christiansen, M., Isbilen, E., Revius, K., & Smithson, M. (under review). Measuring children's auditory statistical learning via serial recall. Manuscript obtained via corresponding author *(manuscript obtained via personal communication with Evan Kidd)*.

*Kidd, E., & Kirjavainen, M. (2011). Investigating the contribution of procedural and declarative memory to the acquisition of past tense morphology: Evidence from Finnish. *Language and Cognitive Processes*, *26*(4/5/6), 794–829. https://doi.org/10.1080/01690965.2010.493735

Krogh, L., Vlach, H., & Johnson, S. (2013). Statistical Learning Across Development: Flexible Yet Constrained. *Frontiers in Psychology*, *3*(598). https://doi.org/10.3389/fpsyg.2012.00598

Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing: An Interdisciplinary Journal*, *33*(6), 1557–1589. https://doi.org/https://doi:10.1007/s11145-020-10018-4

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical learning in specific language impairment: A meta-analysis. *Journal of Speech, Language and Hearing Research*, *60*(12), 3474–3486. https://doi.org/10.1044/2017_JSLHR-L-16-0439

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Auditory statistical learning in children: Evidence from an online measure. *Applied Psycholinguistics*, *40*(2), 279–302. https://doi.org/10.1017/S0142716418000577

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2020). Children with developmental language disorder have an auditory verbal statistical learning deficit: Evidence from an online measure. *Language Learning*, *70*(1), 137–178. https://doi.org/10.1111/lang.12373

Leonard, L. B. (2014). *Children with specific language impairment*. MIT Press.

Lum, J., Conti-Ramsden, G., Morgan, A., & Ullman, M. (2014). Procedural learning deficits in specific language impairment (SLI): A meta-analysis of serial reaction time task performance. *Cortex*, *51*(100), 1–10. https://doi.org/10.1016%2Fj.cortex.2013.10.011

Lum, J., & Kidd, E. (2012). An examination of the associations among multiple memory systems, past tense, and vocabulary in typically developing 5-year-old children. *Journal of Speech, Language and Hearing Research*, *55*(4), 989–1006. https://doi.org/10.1044/1092-4388(2011/10-0137)

*Lum, J., Conti-Ramsden, G., Page, D., & Ullman, M. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex*, *48*(9), 1138–1154. https://doi.org/10.1016/j.cortex.2011.06.001

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), 101–111. https://doi.org/10.1016/S0010-0277(01)00157-3

*Mimeau, C., Coleman, M., & Donlan, C. (2016). The role of procedural memory in grammar and numeracy skills. *Journal of Cognitive Psychology*, *28*(8), 899–908. https://doi.org/10.1080/20445911.2016.1223082

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child speech. *Cognition*, *90*(1), 91–117. https://doi.org/10.1016/S0010-0277(03)00140-9

Misyak, J., & Christiansen, M. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302–331. https://doi.org/10.1111/j.1467-9922.2010.00626.x

Misyak, J., Christiansen, M., & Tomblin, J. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, *1*(31), 1–9. https://doi.org/10.3389/fpsyg.2010.00031

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PloS Medicine*, *6*(7), e1000097. https://doi:10.7326/0003-4819-151-4-200908180-00135

Montgomery, J., Gillam, R., Evans, J., & Sergeev, A. (2017). "Whatdunit?" Sentence comprehension abilities of children with SLI: Sensitivity to word order in canonical and noncanonical structures. *Journal of Speech, Language and Hearing Research*, *60*(9), 2603–2618. https://doi.org/10.1044/2017_JSLHR-L-17-0025

Nissen, M., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1–32. https://doi.org/10.1016/0010-0285(87)90002-8

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

*Obeid, R. (2017). *Exploring the relationship between sequence learning, motor coordination, and language development* (Doctoral dissertation submitted to the graduate faculty in psychology), The city University of New York.

Obeid, R., Brooks, P., Powers, K., Gillespie-Lynch, K., & Lum, J. (2016). Statistical learning in specific language impairment and autism spectrum disorder: A meta- analysis. *Frontiers in Psychology*, *7*(1245). https://doi.org/10.3389/fpsyg.2016.01245

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

*Park, J., Miller, C., Rosenbaum, D., Sanjeevan, T., van Hell, J., Weiss, D., & Mainela-Arnold, E. (2018). Bilingualism and procedural learning in typically developing children and children with language impairment. *Journal of Speech, Language, and Hearing Research*, *61*(3), 634–644. https://doi.org/10.1044/2017_JSLHR-L-16-0409

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language and Communication Disorders*, *50*(1), 106–118. https://doi.org/10.1111/1460-6984.12126

R Core Team. (2018). *R: A language and environment for statistical computing [Computer software]*. R Foundation for Statistical Computing. https://www.r-project.org

Raven, J., Raven, J., & Court, J. (2003). *Manual for Raven's progressive matrices and vocabulary scales [measurement instrument]*. Harcourt.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia*, *67*(2), 1–16. https://doi.org/10.1007/s11881-016-0136-0

Semel, E., Wiig, E., & Secord, W. (2010). *Clinical evaluation of language fundamentals: Dutch version* W. Kort, E. Compaan, M. Schittekatte, & P. Dekker, Trans(Eds.), [Measurement instrument] (3rd ed.). Pearson.

Siegelman, N., Bogaerts, L., Christiansen, M., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B*, *372*(20160059), 20160059. https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavioral Research*, *49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. https://doi.org/10.1016%2Fj.cognition.2007.06.010

Sociaal en Cultureel Planbureau. (2017, February). Statusscores 2016 *[report from social and cultural planning]*. Sociaal en Cultureel Planbureau.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Treiman, R. (2018). Statistical learning and spelling. *Language, Speech and Hearing Services in Schools*, *49*(3S), 644–652. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0122

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analysis: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665. https://doi.org/10.1177/1745691614552498

Ullman, M. (2014). The declarative/procedural model: A neurobiologically-motivated theory of first and second language. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (2nd ed., pp. 135–158). Routledge.

Ullman, M., & Pierpont, E. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex*, *41*(3), 399–433. https://doi.org/10.1016/S0010-9452(08)70276-4

Ullman, M., & Pullman, M. (2015). A compensatory role for declarative memory in neurodevelopmental disorders. *Neuroscience and Biobehavioral Reviews*, *51*, 205–222. https://doi.org/10.1016/j.neubiorev.2015.01.008

van den Bos, K., Spelberg, L., Scheepstra, A., & de Vries, J. (1994). *Klepel [nonce word reading; Measurement instrument]*. Pearson.

van Witteloostuijn, M. (2020). *Examining the contribution of statistical learning to grammar and literacy acquisition: A study of Dutch children with and without dyslexia* (Doctoral dissertation submitted to the University of Amsterdam): LOT Publications.

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2017). Visual artificial grammar learning in dyslexia: A meta-analysis. *Research in Developmental Disabilities*, *70*, 126–137. https://doi.org/10.1016/j.ridd.2017.09.006

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software*, *36* (3), 1048. http://www.jstatsoft.org/v36/i03/.

West, G., Shanks, D., & Hulme, C. (2020). Sustained attention, not procedural learning, is a predictor of language, reading and arithmetic skills in children. *Scientific Studies of Reading*, 1–17. Advance online publication. https://doi.org/10.1080/10888438.2020.1750618

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, *21*(2), 1–13. https://doi.org/10.1111/desc.12552

Wijnen, F. (2013). Acquisition of linguistic categories: Cross-domain convergences. In J. J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech and language: Exploring the evolution of mind and brain*. The MIT press.

Wilson, B., Spierings, M., Ravignani, A., Mueller, J., Mintz, T., Wijnen, F., . . . Rey, A. (2018). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science*. Advance online publication. https://doi.org/10.1111/tops.12381

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Zwart, F., Vissers, C., Kessels, R., & Maes, J. (2019). Procedural learning across the lifespan: A systematic review with implications for atypical development. *Journal of Neuropsychology*, *13*(2), 149–182. https://doi.org/10.1111/jnp.1213