UvA-DARE (Digital Academic Repository)

Children With Developmental Language Disorder Have an Auditory Verbal Statistical Learning Deficit: Evidence From an Online Measure

Lammertink, I.; Boersma, P.; Wijnen, F.; Rispens, J.

Citation for published version (APA):

EMPIRICAL STUDY

Children With Developmental Language Disorder Have an Auditory Verbal Statistical Learning Deficit: Evidence From an Online Measure

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

[a]University of Amsterdam and [b]Utrecht University

Successful language use requires the ability to process nonadjacent dependencies (NADs) that occur in linguistic input. Learning such structural regularities seems

therefore crucial for children, and researchers have indeed proposed that language problems in children with developmental language disorder (DLD), especially problems with grammar, are due to their decreased sensitivity to NADs. Because the evidence supporting this claim is scarce, we compared children with DLD ($n = 36$; $M_{age} = 9.1$ years) and without DLD ($n = 36$; $M_{age} = 9.1$ years) performing a learning task with NADs. Using response times as an online measure of learning NADs, we observed that participants with DLD were less sensitive to NADs than were typically developing peers. The confidence intervals of the effect, however, indicated that the effect was probably small in size. We discuss clinical and theoretical implications of the present study in light of this effect size.

Keywords developmental language disorder; specific language impairment; individual differences; nonadjacent dependencies; statistical learning

Introduction

Children with developmental language disorder (DLD) have problems with language that significantly impact their social interactions and educational progress (Bishop, Snowling, Thompson, & Greenhalgh, 2017). Children with DLD often exhibit difficulties across multiple language areas, and these problems frequently co-occur with deficiencies in other cognitive domains such as attention, working memory, and procedural memory (e.g., Ebert & Kohnert, 2011; Montgomery, Evans, & Gillam, 2018; Ullman & Pierpont, 2005). Even though DLD is a heterogeneous disorder (Bishop et al., 2017), difficulties with learning morphosyntactic and morphological rules are a clinical marker of the disorder. More specifically, correct use of morphemes that mark tense and agreement is notoriously difficult for these children (e.g., see a meta-analysis on past tense production in children with and without DLD by Krok & Leonard, 2015).

Because the core deficit of the language disorder is still unknown (Bishop et al., 2017), theories of its origin keep emerging. Recently, researchers have proposed that children with DLD have a statistical learning deficit, meaning that they are less sensitive to (statistical) regularities in their (linguistic) input (Evans, Saffran, & Robe-Torres, 2009; Hsu & Bishop, 2014a; Lammertink, Boersma, Wijnen, & Rispens, 2017; Obeid, Brooks, Powers, Gillespie-Lynch, & Lum 2016; Wijnen, 2013). Detecting and extracting regularities (statistical patterns) are thought to be fundamental for the earliest stages of language development (Evans et al., 2009), and therefore, it is not surprising that deficits in the ability to detect statistical patterns have been put forward as an explanation for DLD. Yet, in most studies where researchers have investigated statistical learning in DLD, they have focused on statistical learning in the visuomotor

domain (for a meta-analytic overview, see Lum, Conti-Ramsden, Morgan, & Ullman, 2014), on statistical learning at the word segmentation level (e.g., Evans et al., 2009; Haebig, Saffran, & Weismer, 2017; Mayor-Dubois, Zesiger, Van der Linden, & Roulet-Perez, 2014), or on auditory verbal statistical learning in adolescents (Grunow, Spaulding, Gómez, & Plante, 2006; Hsu, Tomblin, & Christiansen, 2014). In most of these studies, researchers did not report on the learning of nonadjacent dependencies (NADs)—a central feature of syntactic processing (Wilson et al., 2018). Therefore, in the present study, we compared NAD learning by children with and without DLD to investigate auditory verbal statistical learning in children with and without DLD.

Background Literature
Nonadjacent Dependency Learning
In classical NAD learning experiments, researchers have auditorily exposed participants to strings of pseudowords in an artificial language. Unbeknownst to the participants, the strings in the language follow a statistical pattern: They consist of three pseudowords (e.g., *tep wadim lut*, *sot wadim mip*), and there is a NAD rule governing the relationship of the first element (*tep* or *sot*) and the last element (*lut* or *mip*), such that the first element predicts the occurrence of the third element (i.e., the co-occurrence probability between the first element and third element is 1.0). After a certain period of exposure to the language, participants perform a grammaticality judgment task in which they are tested with strings that either conform to the NAD rules (e.g., *tep wadim lut*) or violate the NAD rules (e.g., **sot wadim lut*, where the asterisk indicates a violation of the rule). Participants are asked to indicate whether the string with which they are presented follows the same pattern as the strings in the exposure phase or follows a different pattern. If participants are sensitive to the NAD rules, they should endorse strings that conform to the NAD rules more frequently than strings that violate the NAD rules, and thus their correctness probabilities should exceed chance level (Gómez, 2002).

Statistical Learning and Its Relation to Language Proficiency
Researchers have found a link between statistical learning and language proficiency in studies where they have compared statistical learning performance in people with language learning disabilities to statistical learning performance in people without such disabilities. Three meta-analyses have reported a statistical learning deficit in people with DLD (Lammertink et al., 2017; Lum et al., 2014; Obeid et al., 2016). From these meta-analyses (and additional studies published subsequently), it became clear that, although there were ample studies on

statistical learning of children with DLD in the visuomotor domain (approximately 22), there were fewer studies on auditory statistical learning in this group of children (four studies) and that there was only one (recently published) study on auditory NAD learning (reported as specific co-occurrence probability) in children with DLD (Iao, Ng, Wong, & Lee, 2017). Researchers in three of the four studies of auditory statistical learning in children with and without DLD assessed children's sensitivity to statistical structure at the word segmentation level (Evans et al., 2009; Haebig et al., 2017; Mayor-Dubois et al., 2014). In these studies, participating children listened to a continuous stream of auditorily presented syllables in which the transitional probability between adjacent syllables within words was higher (1.0) than the transitional probability between adjacent syllables that spanned word boundaries (e.g., .33). Sensitivity to these differences in transitional probability guided the participants in extracting words from the continuous speech stream. In all three studies, the children with DLD were less sensitive to the differences in transitional probabilities than the typically developing children.

In the fourth study, Lukács and Kemény (2014) used an artificial grammar learning experiment to assess differences in the ability to extract regularities from auditory sequences between children with and without DLD. The researchers constructed the regularities in the auditory sequences to follow different rules, with varying patterns of transitional probability (at the adjacent and nonadjacent level) and with sequences defined at the level of categories instead of at the level of items. As they had hypothesized, Lukács and Kemény found that a significantly smaller proportion of the participating children with DLD showed evidence of learning the rules compared to that of the typically developing children. Finally, Iao et al. (2017) investigated auditory NAD learning in children with DLD and in those without DLD and observed that, when using an offline measure of learning, the children with DLD were less sensitive to NADs than the typically developing children. Taken together, although there has been some work on auditory statistical learning in children with DLD, there have been only two studies in which researches have investigated this type of learning with designs that modeled the acquisition of grammatical structures (Iao et al., 2017; Lukács & Kemény, 2014). Of these two studies, only Iao et al. (2017) investigated children's sensitivity to NAD structures specifically. Given that children with DLD mainly exhibit language difficulties that manifest themselves with NAD structures such as subject–verb agreement and past tense inflection, we deemed it important to further investigate children's sensitivity to this specific co-occurrence probability. In a design different from Iao et al.'s (2017), we assessed children's sensitivity to NADs using both an online and

an offline measure of learning instead of using an offline measure only. In the next section, we discuss how and why it is important that the present study complemented this work by using an online measure of NAD learning.

Another source of evidence for a link between statistical learning and language proficiency has been found in studies showing that individual differences among adults without language learning disabilities while they performed a NAD learning task predicted their comprehension and processing of dependencies in relative clause sentences (Misyak & Christiansen, 2012; Misyak, Christiansen, & Tomblin, 2010). In these studies, adults were asked to read sentences containing relative clauses like "the reporter that attacked the senator admitted the error." Participating adults' processing time measured through a self-paced reading task (Misyak et al., 2010) and their understanding of these sentences (Misyak & Christiansen, 2012) correlated with their performance on an online NAD learning task (Misyak et al., 2010) and an offline NAD learning task (Misyak & Christiansen, 2012). The fact that these adults needed to track the NAD between the head noun *reporter* and main verb *admitted* in order to understand the sentence might have explained these correlations. To the best of our knowledge, in no studies have researchers investigated the specific links between NAD learning and primary-school-aged children's understanding and/or processing of relative clause sentences. There may be two explanations for this. First, there have been only two (published) studies on NAD learning in primary-school-aged children (Iao et al., 2017; Lammertink, van Witteloostuijn, Boersma, Wijnen, & Rispens, 2018). Both these studies evaluated NAD learning in children but did not correlate children's individual NAD learning performance to an individual measure of relative clause sentence processing and/or understanding. And second, it takes children a relatively long period of time to understand and correctly use relative clause structures (for an overview, see Duinmeijer, 2016). Spit and Rispens (2018) used relative clause constructions to investigate the relationship between visuomotor statistical learning, measured through a serial reaction time task (Nissen & Bullemer, 1987), and syntactic proficiency in gifted primary-school-aged children and their typically developing peers. Even though the gifted children scored better on the relative clause comprehension task than their typically developing peers, Spit and Rispens found no evidence for or against a relationship between visuomotor statistical learning and children's relative clause sentence understanding.

Relative clause constructions are not the only linguistic structure governed by NADs. NADs are also present in other morphological and morphosyntactic constructions such as subject–verb agreement, plural nouns, and the past

tense. Many subtests of standardized language test batteries assess, among other grammatical structures, children's production and understanding of these constructions. In a recent meta-analysis, Hamrick, Lum, and Ullman (2017) reported a statistically significant positive correlation between performance on a serial reaction time task and (morpho)syntactic production and comprehension tasks from standardized language test batteries: Test for the Reception of Grammar (Bishop 2003), Épreuve de compréhension syntaxico-sémantique: Adaptation française du TROG: Reception of Grammar Test (Lecocq, 1998), Évaluation du langage oral (Khomsi, 2001), Batterie langage oral, langage écrit, mémoire, attention (Chevrie-Muller, Maillart, Simon, & Fournier, 2010), and Action Picture Test (Renfrew, 2003) in typically developing children. The same link has recently been investigated in a meta-analysis combining children with DLD and without DLD (Lammertink, Boersma, Wijnen, & Rispens, 2019a). In this meta-analysis, Lammertink and colleagues found no evidence for or against a correlation between serial reaction time performance and expressive grammar knowledge in the pooled group of children. This may not be surprising given that most studies on the relationship between serial reaction time performance and grammar knowledge in children with DLD reported statistically nonsignificant (both positive and negative) correlations: positive (Gabriel, Maillart, Guillaume, Stefaniak, & Meulemans, 2011; Gabriel, Stefaniak, Maillart, Schmitz, & Meulemans, 2012; Lum, Conti-Ramsden, Page, & Ullman, 2012) and negative (Desmottes, Meulemans, & Maillart, 2016; Gabriel, Meulemans, Parisse, & Maillart, 2015). Interestingly, Lammertink et al. also found no evidence that the strength of the relationship between serial reaction time task performance and expressive grammar knowledge differs between children with and without DLD.

Statistical Learning and Its Methodological Challenges

Researchers have raised concerns regarding the interpretability of the outcome measure of the design used in classical statistical learning experiments (Siegelman, Bogaerts, & Frost, 2017). A first concern has been that metalinguistic skills or explicit knowledge might have influenced the judgment measure. If indeed performance depends on metalinguistic skills, this impedes valid assessment of children's learning in a NAD task because children acquire metalinguistic skills relatively late (Bialystok, 1986). Also, the acquisition of metalinguistic knowledge may rely more on rote learning strategies rather than on statistical learning (or rule learning) strategies. A second concern had been that children tend to accept all strings, and thus they often show a yes bias when they are asked to make judgments (Ambridge & Lieven, 2011). Because

an increasing number of researchers have stressed the importance of measuring statistical learning in a different way than through grammaticality judgments, several novel measures have been proposed. Following this trend, we decided to use response times as an online measure of NAD learning, in particular measuring the disruption peak that occurs in the response time pattern when items are presented that are discordant with NAD rules. Previous work has shown that disruption peaks reflect sensitivity to NADs in adults (López-Barroso, Cucurell, Rodríguez-Fornells, & de Diego-Balaguer, 2016; Misyak et al., 2010; Vuong, Meyer, & Christiansen, 2015) and in primary-school-aged children (Lammertink, van Witteloostuijn et al., 2018). The use of disruption peaks as an index of statistical learning has its roots in the serial reaction time task literature (Nissen & Bullemer, 1987), and the reason to work with disruption peaks rather than a decrease in response times over the first few training blocks is that such a response time decrease is not necessarily the result of statistical learning. The decrease may also arise as a consequence of practice, which makes it difficult to disentangle statistical learning from motor or cue learning (Kidd & Kirjavainen, 2011, but see Kuppuraj, Duta, Thompson, & Bishop, 2018, for a potential solution to this problem).

Despite our concerns about the interpretability of the offline measures of statistical learning, we measured participants' behavior in an offline forced-choice task as well. Response times are not necessarily a substitute for the judgment measure. It could for instance be that the online reaction time measure and the offline judgment measure tap into different representations of acquired knowledge or that they are sensitive to different learning strategies (see also Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015; Isbilen, McCauley, Kidd, & Christiansen, 2017; Misyak et al., 2010).

The Present Study

To summarize, the aim of the present study was to investigate auditory verbal statistical learning of NADs in children with and without DLD. Our confirmatory research question tested the hypothesis that children with DLD are less sensitive to NADs than their typically developing peers; hence, we expected children with DLD to show a statistical learning deficit. We evaluated NAD learning in both groups of children through an online measure in which the size of a disruption peak in response times was used as an estimate of children's sensitivity to the NADs. We predicted that children with DLD would have an auditory verbal statistical learning deficit if their disruption peak was smaller than the disruption peak observed in their typically developing peers. As explained later, we used the interaction between the group variable and

the predictor variable that estimated the size of the disruption peak to answer our confirmatory research question. Because we used verbal material in the auditory domain in our tasks, we expected that verbal short-term memory (Hsu & Bishop, 2011) and verbal working memory (Misyak & Christiansen, 2012; Wilson et al., 2018) might also play a role in participants' successful detection of the NAD rules. We therefore controlled for these measures in our statistical model.

Besides our confirmatory research question, we also used data from the present study to explore four additional questions. First, one anonymous reviewer asked us to explore whether the difference in participants' response times between the first training block and the last training block (third block) was larger for typically developing children than for children with DLD, and second, whether the difference in response times between this first training block and the last training block correlated with the size of children's disruption peak. Third, because we investigated differences in online NAD learning between children with and without DLD (confirmatory research question), we also explored more specifically the association between NAD learning and two tasks that measured children's knowledge of grammatical rules in the expressive domain. Finally, given the abovementioned methodological considerations regarding the use of offline measures of statistical learning, we had some concerns as to whether we could assess NAD learning through an offline measure; this was explored by evaluating children's behavior in an offline forced-choice task.

Method

Participants

We recruited 37 children with DLD and 59 typically developing children aged between 7 and 11 years to participate in our study.[1] At the end of the study, we excluded one participant with DLD and five typically developing participants. The final sample included 36 children with DLD (8 females, 28 males) and 36 typically developing children (9 females, 27 males). 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. Thus, we recruited and tested as many children as possible in the available recruitment time. We nevertheless expected the power of the experiment to detect a medium-sized effect to be guaranteed because the number of participants per group (36) was large for this type of study (see Discussion section). The widths of the resulting confidence intervals would reveal whether this expectation was warranted.

Table 1 Summary of participants' characteristics by group

Characteristics		DLD ($n = 36$)	TD ($n = 36$)
Age (months)	M	109	109
	Range	94–125	93–125
Nonverbal intelligence[a]			
Raw	M	36	36
	Range	23–49	26–55
Standardized (percentiles)	M	63	64
	Range	17–96	20–98
Social economic status[b]	M	0.22	−0.06
	Range	−2.57–2.09	−1.28–1.15

Note. DLD = developmental language disorder; TD = typically developing. [a]Raven Colored Progressive Matrices subtest of Raven's Progressive Matrices and Vocabulary Scales (Raven et al., 2003). [b]Based on data from *Statusscores 2016* (Sociaal en Cultureel Planbureau, 2017).

We obtained ethical 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. Table 1 provides details of participants' age, nonverbal intelligence, and socioeconomic status. We derived their socioeconomic status from a combined score that took the mean education level, mean income, and mean working status of the people living in a particular district (defined per zip code) into account (Sociaal en Cultureel Planbureau, 2017). This score has a Dutch average of 0, and the higher the score, the higher the socioeconomic status. We based the socioeconomic status of the participants with DLD on either their home address ($n = 22$) or school address ($n = 14$). We based the socioeconomic status of the typically developing participants on their school address (four different schools across the Netherlands).

Recruitment and Inclusion of Children With Developmental Language Disorder

We recruited the participating children with DLD through four national organizations in the Netherlands (Royal Dutch Auris Group, Royal Dutch Kentalis, Viertaal, and Pento), through an association for parents of children with DLD (Stichting Hoormij), and through self-employed speech therapists. All

participants in this group had been diagnosed with DLD by licensed clinicians and met the following criteria: (a) they had scored 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) none had been diagnosed with autism spectrum disorder, attention deficit hyperactivity disorder, or with other (neuro)physiological problems. Finally, our test battery included the Raven Colored Progressive Matrices subtest (Raven, Raven, & Court, 2003), a standardized measure of nonlinguistic intelligence, on which the participants had to obtain a percentile score of at least 17 to be included in our final sample. A percentile score of 17% was the lower bound of the normal range, and therefore, if participants had a percentile score below 17%, they were assessed as having below average nonverbal intelligence. At the time that we started recruitment for this project, children with language difficulties had to have a nonverbal intelligence score of at least average to get a diagnosis of specific language impairment/DLD in the Netherlands. This was also why we decided to include only children who met this IQ criterion (and thus a Raven Colored Progressive Matrices score of at least 17%). Only shortly thereafter, Bishop et al. (2017) made their recommendation that low nonverbal ability should not preclude a diagnosis of DLD. At the end of the study, we excluded one participant with DLD because of an only recently diagnosed hearing problem.

Recruitment and Inclusion of Typically Developing Children
We recruited the typically developing children from four different primary schools across the Netherlands. Because these typically developing children had never taken a standardized language assessment test battery prior to participating in the present study, we used their scores on the Raven Colored Progressive Matrices subtest (Raven et al., 2003) and a subset of the language tasks (see below) that were administered as part of our own test battery as inclusion criteria. We excluded five typically developing children because they scored below the normal range on the Raven Colored Progressive Matrices subtest and/or they scored below the normal range on two or more of the following language tasks: the Een-Minuut-Test, a one-minute real-word reading test (Brus & Voeten, 1979); the Klepel, a two-minute nonce word reading test (van den Bos, Spelberg, Scheepstra, & de Vries, 1994); the Schoolvaardigheidstoets Spelling, a test of spelling (Braams & de Vos, 2015); and/or the Clinical Evaluation of Language Fundamentals–Dutch version (Semel et al., 2010), a test of

sentence recall. The normal range included scores from 1 standard deviation below the standardized mean (norm scores: $M = 10$; percentiles: $M = 50\%$) to scores 1 standard deviation above the standardized mean, thus extending between 8 and 12 (norm scores) or between 17% and 86% (percentiles). Additionally, we excluded one typically developing participant because this child was diagnosed with attention deficit hyperactivity disorder. From the remaining 53 typically developing children, we selected 36 participants who matched best our DLD sample, taking age (maximum age difference of three months), gender, socioeconomic status, and nonverbal intelligence into account.

Materials
Measure of Statistical Learning
We used a NAD learning task to measure participants' sensitivity to statistical structure in an artificial language (see Lammertink, van Witteloostuijn et al., 2018, for an elaborate description of this task, and see López-Barroso et al., 2016, for its original adult version). Disruption in response times (i.e., slower response times to items in which NAD rules are disrupted compared to items that satisfy NAD rules) served as our measure of participants' sensitivity to the NADs. We presented the NAD task on a Microsoft Surface 3 tablet computer using the E-prime software (Version 2.0; 2012). We recorded response times with an external button box attached to the computer. We played the auditory stimuli to the participants over Sennheiser HD 201 headphones.

During the online part of the NAD task, we exposed the participants to three-element utterances of an artificial language and asked them to press either a green button if the third element that they heard was a specific target (e.g., *lut*) or a red button if the third element was not this specific target (see Figure 1). In all utterances, Element 1 was a monosyllabic Dutch nonce word (e.g., *tep*), Element 2 was a bisyllabic Dutch nonce word (e.g., *wadim*), and Element 3 was again a monosyllabic Dutch nonce word (e.g., *lut*). We divided the utterances into three trial types. Two types comprised a NAD between Element 1 and Element 3: *tep* X *lut* or *sot* X *mip.* In these examples, X indicated the bisyllabic element that was drawn from a pool of 24 different elements (see Table 2 for the list of elements) following Gómez (2002). There were two versions of the experiment with either *lut* (Version 1) or *mip* (Version 2) as the target word. We randomly assigned participants to one of the two versions. We divided the NAD types into target trials ending with the target word (Version 1: *lut*; Version 2: *mip*), which thus required participants to press a green button, and nontarget trials ending with the nontarget word (Version 1: *mip*; Version 2: *lut*), which thus required participants to press a red button. The third type were filler

Figure 1 Example of the online nonadjacent dependency task. [Color figure can be viewed at wileyonlinelibrary.com]

Table 2 Overview of the 24 X elements and 24 F elements used to build the target items, nontarget items, and filler items

X elements	F elements
banip, biespa, dapni, densim, domo, fidang, filka, hiftam, kasi, kengel, kubog, loga, movig, mulon, naspu, nilbo, palti, pitok, plizet, rasek, seetat, tifli, valdo, wadim	bap, bif, bug, dos, dul, fas, fef, gak, gom, hog, huf, jal, jik, keg, ket, kof, naf, nit, nup, pem, ves, wop, zim, zuk

trials, which did not contain a NAD (and no *lut* or *mip*), and therefore they always required participants to press a red button.

The experiment consisted of five blocks. Four of these blocks (Training Block 1, Training Block 2, Training Block 3, and a fifth recovery block) contained target trials and nontarget trials with the NAD rules, as we described above (i.e., NAD blocks). In these blocks, the third element of the target trials

and nontarget trials could thus be predicted from the first element. The fourth block (disruption block) was exceptional: It contained target trials and nontarget trials in which the dependency between the first and third elements was disrupted, that is, the target element or nontarget element (*lut* or *mip*) was now preceded by a variable filler element (F element), that is, never *tep* or *sot*, in the first position. In these trials, the third element of the target trials and nontarget trials could thus no longer be predicted from the first element. If participants were sensitive to the NADs, we predicted that their response times to target trials and nontarget trials in the disruption block would be slower than their response times to these items in the third training block and in the recovery block. We refer to this difference in response times as the *disruption peak*. All NAD blocks contained 24 target trials (i.e., *tep* X *lut* in Version 1), 24 nontarget trials (i.e., *sot* X *mip* in Version 1), and 12 filler trials (i.e., no NAD and ending in something other than *lut* or *mip*). The disruption block contained 12 target trials (i.e., no NAD, but *lut* final in Version 1), 12 nontarget trials (i.e., no NAD, but *mip* final in Version 1), and six filler trials (i.e., no NAD and ending in something other than *lut* or *mip*).

After completing these five blocks, participants received instructions for the offline forced-choice task. We told them that they would hear an utterance and that they had to decide whether they had heard this utterance previously. We presented participants with 18 utterances; two of these utterances had a completely different structure from the utterances in the online phase (**kasi kubog kengel* and **banip dapni nilbo*) and served as control items. The remaining 16 utterances were actual test items. These test items consisted of four types: (a) correct NAD items with familiar X elements (*tep palti lut*; *sot densim mip*; *tep hiftam lut*; *sot fidang mip*), (b) incorrect NAD items with familiar X elements (*sot filka lut; *tep loga mip; *sot plizet lut; *tep rasek mip), (c) correct NAD items with novel X elements (*tep sulep lut*; *sot dieta mip*; *tep nukse lut*; *sot noeba mip*), and (d) incorrect NAD items with novel X elements (*sot rolgo lut; *tep gopem mip; *sot wiffel lut; *tep dufo mip). The familiar X elements were eight of the 24 X elements that the participants had already heard during the exposure phase (*palti*, *densim*, *hiftam*, *fidang*, *filka*, *loga*, *plizet*, *rasek*; see Table 2). The two item types with novel X elements contained eight novel X elements (*sulep*, *dieta*, *nukse*, *noeba*, *rolgo*, *gopem*, *wiffel*, *dufo*). We added these items to test for generalization of the rule. The participants had to declare verbally whether they had heard the utterance previously, and the experimenter recorded their responses in E-prime. In total, the experiment took approximately 30 minutes: 20 minutes for the online phase; 5 minutes for the offline phase; and 5 minutes for instructions, practice, and pauses.

Measures of Morphosyntax and Morphology

We administered two measures to tap into participants' expressive knowledge of grammatical rules: the Sentence Recall task and the Word Structure task from the Clinical Evaluation of Language Fundamentals–Dutch version (Semel et al., 2010). We used the Sentence Recall task as an index of participants' morphosyntactic knowledge. In this task, we asked participants to repeat sentences with increasing length and complexity. Following the guidelines of the Clinical Evaluation of Language Fundamentals–Dutch version, we assigned points to responses based on the number of errors that participants made in the recalled sentence, with 3 points for fully correct repetitions, 2 points for repetitions with one error, 1 point for repetitions with two or three errors, and 0 points for repetitions with four or more errors. The task terminated when participants scored 0 points on five consecutive sentences. The maximum number of points that participants could obtain was 93.

We assessed participants' morphological knowledge at the word level with the Word Structure task. In this task, we orally presented participants with 30 incomplete sentences that described a picture and asked participants to complete the sentences. Missing words were either plurals, pronouns, inflectional morphemes, derivational morphemes, or comparatives. We awarded 1 point for each correct completion, with a maximum total of 30 points.

Other Cognitive and Language Measures

We also collected measures of participants' nonverbal intelligence (Raven et al., 2003), receptive vocabulary size (Peabody Picture Vocabulary Task-III-NL; Schlichting, 2005), verbal short-term memory (Digit Span Forward; Semel et al., 2010), verbal working memory (Digit Span Backward; Semel et al., 2010), and sustained attention (Tel mee! subtest from the Test of Everyday Attention for Children; Manly, Robertson, Anderson, & Nimmo-Smith, 2010). Table 3 provides a short description of each measure.

Procedure

The present study was part of a larger research project about the relationship between statistical learning and grammar and literacy acquisition in children with and without DLD, and therefore, the total task battery contained more tasks than we have reported here. All children who participated in the present study completed this full battery, which took two to four sessions (each lasting approximately 1 hour), spread over 2 to 3 weeks for each child. Each test session started with a statistical learning task—the NAD learning task, a visual statistical learning task, or a serial reaction time task—and was then followed

Table 3 Description of the measures used in the study

Task	Description	Possible range (raw scores)
Raven's Progressive Matrices and Vocabulary Scales (Raven et al., 2003)	*Nonverbal intelligence* Children are asked to complete a visual pattern by selecting the correct missing pattern from six or eight possible options.	1–60
Peabody Picture Vocabulary Test-III-NL (Schlichting, 2005)	*Receptive vocabulary size* Children hear a word a have to choose the correct referent out of four pictures.	1–204
Digit Span Forward from the Clinical Evaluation of Language Fundamentals (Semel et al., 2010)	*Verbal short-term memory* Children are asked to immediately repeat a number of sequences of increasing length in the same order.	0–16
Digit Span Backward from the Clinical Evaluation of Language Fundamentals (Semel et al., 2010)	*Verbal working memory* Children are asked to immediately repeat a number of sequences of increasing length in reversed order.	0–14
Tel Mee! From the Test of Everyday Attention for Children (Manly et al., 2010)	*Sustained attention* Children are asked to count sounds. Each trial has a different number of sounds to count (ranging from 9 sounds to 14 sounds). The pauses between the sounds in each trial are of variable length.	0–10

by a set of cognitive and language measures. Participants completed the verbal short-term memory task and verbal working memory task in the same session as they did the NAD learning task. They completed the Sentence Recall task, Word Structure task, sustained attention task, and the Raven Colored Progressive Matrices subtest in the session with the serial reaction time task, and finally, they completed the Peabody Picture Vocabulary Test-III-NL task in the session with the visual statistical learning task. We counterbalanced the order in which participants performed the different sessions. The results for the other

statistical learning tasks are reported in Lammertink, Boersma, Wijnen, and Rispens (2018) and Lammertink, Boersma, Wijnen, and Rispens (2019b). For the typically developing participants, we collected the data in a quiet room at their schools. We collected data for the participants with DLD either in a quiet room in their schools ($n = 22$) or in their homes ($n = 14$).

Data Analysis

We have provided all data and scripts (including full model outcomes) used in the analyses through the Open Science Framework (https://osf.io/8a3yv). During the online part of the statistical learning task, we recorded both participants' accuracy and response times. For our confirmatory analysis, we selected participants' correct responses to target and nontarget items only in the third training block, the disruption block, and the recovery block. We measured response times in milliseconds from the onset of the target item or the nontarget item. For analysis, we normalized the raw response times to make the data satisfy more closely the assumption of normally distributed model residuals, which is a central assumption of linear mixed-effects model analysis. We used package lme4 (Version 1.1.17; Bates, Maechler, Bolker, & Walker, 2015) for the R programming language (R Core Team, 2018) to conduct the analyses. The advantage of working with transformed response time data (in general) over excluding outlier observations in order to satisfy model assumptions is that one can include all observations and does not have to apply an arbitrary criterion, which can vary enormously between studies, for removing observations (Simmons, Nelson, & Simonsohn, 2011). Visual inspection of the model residuals from our raw response time model and normalized response time model indeed indicated that the residuals of the model with normalized response times were more symmetrically distributed than the residuals of the model with raw response times (see histograms at https://osf.io/8a3yv). Therefore, we decided to continue working with normalized response times.

We normalized the response time data with a rank-order transformation. We could not apply the commonly used log-transformation because participants' response times could be negative (i.e., if a participant had learned to predict the third word from the first word and thus pressed the button before the onset of the third word). In transforming the observations, we first sorted all K raw reaction time observations in ascending order, then assigned each ranked observation a ranking number r (from 1 to K; Baguley, 2012, pp. 254–358). Subsequently, we normalized the ranked observations by replacing each observation by the $(r - 0.5)/K$ quantile of the normal distribution. This

normalization allows researchers to interpret the resulting response time values as optimally distributed z values.

We analyzed these normalized response time data using a linear mixed-effects model that fitted normalized response time as a function of the ternary predictor variable block (the third training, disruption, and recovery blocks), the binary predictor variables group (DLD, typically developing), targetness (nontarget, target), and experiment version (version 1, version 2), and the continuous predictor variables verbal short-term memory performance and verbal working memory performance. We refer to this model as the confirmatory disruption peak model. The confirmatory disruption peak model included the main effects of the predictor variables block, group, targetness, and experiment version, as well as all interactions between these predictors. We included verbal short-term memory performance and verbal working memory performance as main effects and in interaction with only the predictor variables block and group because these were the predictors of interest for our confirmatory analysis. We coded all binary and ternary predictors in the model with orthogonal sum-to-zero contrasts (for the specific contrast settings see Appendix S1 in the Supporting Information online), and we centered the continuous variables and scaled them with the scale function in R (R Core Team, 2018). Finally, the random-effects structure of the confirmatory disruption peak model contained by-subject ($N = 72$) and by-item (X element: $N = 24$) random intercepts, by-subject random slopes for the main effects of block and targetness, and by-item random slopes for the main effects of group and of experiment version.[2] This was the maximal random effects structure justified by the design: It contained by-subject random slopes for the within-subject predictor variable block of our confirmatory research question and by-item random slopes for the between-subject predictor variable group of our confirmatory research question (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2018).

We hypothesized that, if participants were sensitive to the NADs, their normalized response times to target and to nontarget items should show a disruption peak (Lammertink, van Witteloostuijn et al., 2018). Furthermore, if NAD learning is related to language proficiency, then this disruption peak should have been lower (or even nonexistent) in the participants with DLD compared to the typically developing participants. The size of the disruption peak was estimated by the first contrast of the predictor variable block (with the disruption block coded as $+\frac{2}{3}$ and both the third training block and the recovery block coded as $-\frac{1}{3}$). We expected that this predictor in interaction with the predictor group (typically developing coded as $+\frac{1}{2}$ and DLD coded

Table 4 Mean and range values for raw and (when available) standardized scores for the participants' performance in the tasks

Task	DLD ($n = 36$)		TD ($n = 36$)		DLD–TD comparison		
	M	Range	M	Range	t	p	95% CI for M_{diff}
Digit Span Forward (Clinical Evaluation of Language Fundamentals)							
Raw	6.2	3–9	8.9	6–12	–7.7	<.001	[–3.4, –2.0]
Standardized[a]	6[c]	1[c]–12	11	6[c]–15			
Digit Span Backward (Clinical Evaluation of Language Fundamentals)							
Raw	3.3	2–5	4.3	2–8	–3.4	.0011	[–1.6, –0.4]
Standardized[a]	8	4[c]–12	10	5[c]–16			
Tel mee! (Test of Everyday Attention for Children)							
Raw	7.2	1–10	7.6	3–10	–0.8	.44	[–1.4, +0.6]
Standardized[a]	8	1[c]–13	9	3[c]–13			
Peabody Picture Vocabulary Test-III-NL							
Raw	101	78–118	115	98–140	–5.8	<.001	[–18.0, –8.9]
Standardized[b]	33	1[c]–84	63	6[c]–95			
Sentence recall							
Raw	31	12–67	59	32–81	–9.2	<.001	[–35, –22]
Standardized[a]	5[c]	1[c]–13	11	3[c]–16			
Word structure							
Raw	22	12–29	28	22–30	–7.4	<.001	[–8, –4]

Note. DLD = developmental language disorder; TD = typically developing. [a]Norm scores. [b]Percentile scores. [c]Standardized scores that fell below the normal range; the normal range included scores from 1 standard deviation below the standardized mean (norm scores: $M = 10$; percentile scores: $M = 50\%$) to scores 1 standard deviation above the standardized mean, thus ranging from 8 to 12 (norm scores) or from 17% to 86% (percentile scores).

as $-\frac{1}{2}$) would allow us to answer our confirmatory research question. The predictor variables experiment version, verbal short-term memory, and verbal working memory were not of direct interest for our research question, but we included them to control for their potential influence on learning. We decided not to control for sustained attention because we had no evidence that our participants with DLD differed from our typically developing participants on this measure (see Tel mee! results in Table 4). We assessed the statistical significance of the predictors via 95% profile confidence intervals and obtained the corresponding p values from the profiles iteratively (see get.p.value function in R functions script at https://osf.io/8a3yv). Unless we explicitly specify so
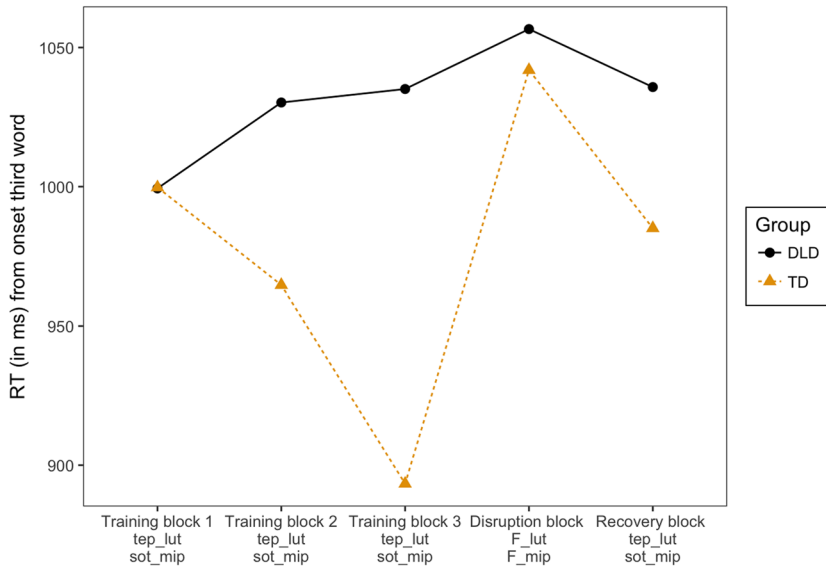
Figure 2 Participants' raw response times (RTs) across all five blocks of the online exposure phase. DLD = developmental language disorder; TD = typically developing. [Color figure can be viewed at wileyonlinelibrary.com]

otherwise, our significance tests assessed whether a value is reliably different from 0.

In addition to our confirmatory research question, we explored four other questions. We cannot draw any confirmatory conclusions from these additional exploratory analyses. First, guided by our descriptive visualization of participants' raw response times across all five blocks of the online exposure phase (see Figure 2), one anonymous reviewer asked us to explore whether the difference between participants' response times in Training Block 1 and their response times in Training Block 3 (i.e., the response time gain) was larger for typically developing participants than for participants with DLD. In exploring this first issue, we analyzed participants' normalized response time data across the first three training blocks with a model that we designated as the exploratory learning speed model and that was very similar to the confirmatory disruption peak model (see above and https://osf.io/8a3yv). The difference was that this model contained data from the first three training blocks (instead of the third training block, disruption block, and recovery block) and thus the ternary predictor variable block was now replaced by the ternary predictor variable training block. Because the effect of interest lay in the size of participants' response

time gain from Training Block 1 to Training Block 3, we set the contrasts of the predictor training block such that a positive estimate of the second contrast of the predictor (with Training Block 1 coded as $+\frac{1}{2}$ and Training Block 3 coded as $-\frac{1}{2}$) estimated this response time gain. We expected that the interaction of the predictor variable response time gain with the predictor variable group, would answer this first exploratory question.

The second question that the anonymous reviewer asked us to explore was whether there was a correlation between participants' response time gain and the size of their disruption peak. In exploring this issue, we first extracted with the ranef function in R (Bates et al., 2015) participants' random slopes for response time gain (from the exploratory learning speed model) and their random slopes for disruption peak (from the confirmatory disruption peak model) and used these random slopes as individual response time gains and individual disruption peaks, respectively. If the individual response time gains were positively correlated with the individual disruption peaks, then this might be a preliminary indication that participants response time gain and their disruption peaks measure similar constructs.

Third, we were also interested in exploring whether there are links between NAD learning and morphosyntax/morphology. We now used the same individual disruption peaks (i.e., random effects of the predictor variable disruption block from the confirmatory disruption peak model) as we had used for the link between participants' response time gain and the size of their disruption peak to explore the link between NAD learning and grammar. We assumed that participants with relative high disruption peaks would be better statistical learners than participants with lower disruption peaks.

Finally, we explored participants' response behavior on the offline forced-choice task in a generalized linear mixed-effects model using package lme4 (Bates et al., 2015). In this model, the dependent variable was endorsement rate. We coded every utterance to which a participant responded positively (i.e., with "yes, I've heard this utterance before") as 1 and every utterance to which a participant responded negatively (i.e., with "no, I've not heard this utterance before") as 0. We fitted endorsement rate as a function of the binary predictor variables generalization (novel, familiar), rule (rule, violation), group (DLD, typically developing), and experiment version (version 1, version 2), and the continuous predictor variables verbal short-term memory and verbal working memory. We included all binary predictors in interaction with each other, and we included the continuous predictors in interaction with only the predictors rule, generalization, and group (the predictors of interest to our research question). The random-effects structure of the offline model contained

by-subject ($N = 72$) and by-item (X-element: $N = 16$) random intercepts, by-subject random slopes for the main effect and interaction of generalization and rule, and by-item random slopes for the main effect and interaction of group and experiment version (Barr et al., 2013, Bates et al., 2018). We coded all binary predictors with orthogonal sum-to-zero contrasts, and we centered and scaled the continuous predictors (for the specific contrast settings, see Appendix S1). We assessed the statistical significance of the predictors using 95% Wald confidence intervals.

Results

Background Measures: Group Comparisons on the Cognitive and Language Tasks

Table 4 presents the raw scores and, when available, the standardized norm or percentile scores for the cognitive and language tasks (described in Table 3) for both groups. Between-group *t* tests (see Table 4) showed that the participants with DLD performed more poorly than the typically developing participants on all cognitive and language tasks except the sustained attention task.

An Online Nonadjacent Dependency Learning Deficit in Developmental Language Disorder

Online Measure: Descriptive Data

A priori we decided to exclude participants from the analysis if their accuracy on the online part of the task was lower than 60% (Lammertink, van Witteloostuijn et al., 2018). Responses were coded as incorrect if participants pressed the wrong button color or if they did not press the button at all. None of the participants had to be removed by this criterion, and we had no evidence that the participants with DLD made more (or fewer) errors than the typically developing participants, pooled over all five blocks and all item types: accuracy for the participants with DLD = 91%; accuracy for the typically developing participants = 94%, $t = -1.59$, $p = .12$, 95% CI of group difference [–0.061%, 0.0069%] (see Data Preprocessing script at https://osf.io/8a3yv). After removing participants' incorrect responses, we plotted their response time trajectory (see Figure 2). We displayed these raw response times only for ease of exposition; they do not represent the outcome of our confirmatory hypothesis testing. Therefore, (descriptive) differences in these raw response times cannot be used to interpret the strength of the effects reported later or to draw any conclusions with respect to our confirmatory research question.

Table 5 Outcome of the linear mixed-effects model for the reaction time data (8,015 observations)

Random effects of subjects ($N = 72$)	SD (Δz)	Correlation Intercept	Disruption peak	Pre–post disruption
Intercept	0.35			
Disruption peak	0.16	–0.31		
Pre–post disruption	0.24	–0.20	0.45	
Targetness	0.11	0.61	–0.32	–0.07

Random effects of X element ($N = 24$)	SD (Δz)	Correlation Intercept	Experiment version	
Intercept	0.31			
Experiment version	0.04	–0.75		
Group	0.07	0.36	0.35	
Residual	0.84			

Fixed effects	B (Δz)	95% CI (Δz)	t	p
Disruption peak × Group[a]	0.19	[0.02, 0.36]	2.23	.03

Note. The full model outcome (including all predictors) can be found in the R markdown script at https://osf.io/8a3yv and an operationalization of the predictors in Appendix S1.
[a]Relevance is confirmatory.

Online Measure: Confirmatory Results

We report only the estimates for the predictors that are relevant for our confirmatory hypothesis testing. The full model outcomes are available at https://osf.io/8a3yv. As we explained previously, we expected that the model estimate for the interaction between the predictor estimating the size of the disruption peak and the predictor variable group would answer our confirmatory research question. The estimate was positive, $\Delta z = 0.19$, $t = 2.23$, 95% profile CI [0.02, 0.36], $p = .03$ (see also Table 5 and Figure 3), which indicated that the disruption peak was between 0.02 and 0.36 standard deviations (of pooled normalized response times) higher in typically developing children than in children with DLD. To obtain an estimate for the range of standardized effect sizes that might be reliably detected, we divided the lower and upper bound of the confidence interval by the residual standard deviation of the model (residual $SD = 0.84$) and observed that the disruption peak was between 0.02 and 0.43 times higher in typically developing children than in children with DLD. Finally, to explore the Group × Disruption Peak interaction, we fitted two additional
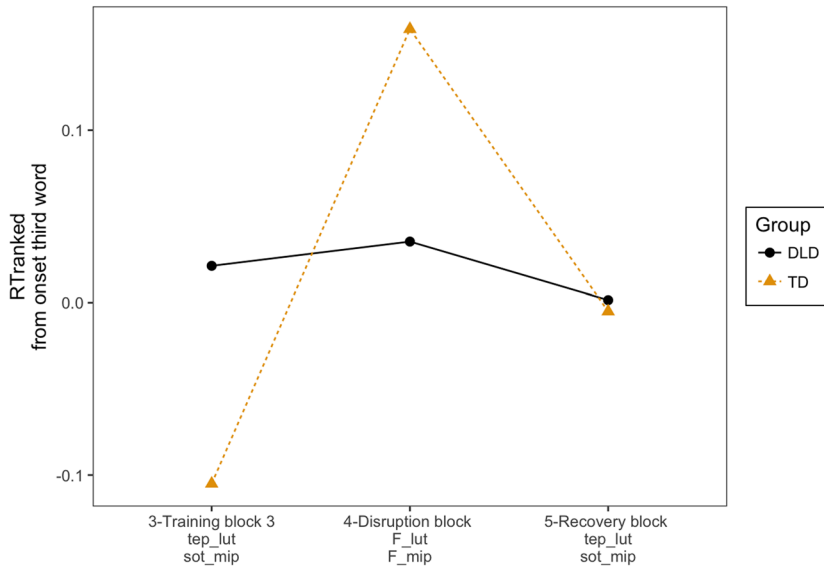
Figure 3 Interaction between the size of the disruption peak and the predictor variable group. RT = reaction time; DLD = developmental language disorder; TD = typically developing. [Color figure can be viewed at wileyonlinelibrary.com]

models in which we re-referenced the contrast coding such that we obtained an estimate for the size of the disruption peak in participants with DLD and in typically developing participants separately. For participants with DLD (with DLD coded as 0, and typically developing as 1), the model estimate for the size of the disruption peak was positive but nonsignificant, $\Delta z = 0.03$, $t = 0.42$, 95% profile CI [–0.10, 0.15], $p = .68$, and therefore we had no evidence that children with DLD were sensitive to the NADs. For typically developing participants (with typically developing coded as 0, and DLD as 1), the estimate for disruption peak was positive and statistically significant, $\Delta z = 0.21$, $t = 3.62$, 95% profile CI [0.09, 0.33], $p < .001$, from which we could conclude that typically developing children were sensitive to the NADs. Taking these results together, we concluded that typically developing children had a positive disruption peak, whereas this disruption peak in children with DLD was lower—if it existed at all—and thus we could speak of a NAD learning deficit in children with DLD.

In addition to providing an estimate for the range of standardized effects sizes for the between-group difference that might be reliably detected,

we also assessed the internal consistency of the online measure (i.e., size of disruption peak). To do so, we computed the split-half reliability: Spearman-Brown corrected Pearson correlation between the size of participants' individual disruption peak for even items (random slopes for the predictor disruption peak from the linear mixed-effects model that included data for even items only) and the size of participants' individual disruption peak for odd items (random slopes for the predictor disruption peak from the linear mixed-effects model that included data for odd items only). The split-half reliability was .79, 95% CI [.66, .87].

Online Measure: Exploratory Results
From the visualization of participants' raw response times across the five blocks (Figure 2), two exploratory questions arose: (a) whether the gain in response time from Training Block 1 to Training Block 3 was larger for typically developing children than for children with DLD and (b) whether this gain in response time was associated with the size of participants' individual disruption peak. To explore the first question, we fitted the exploratory learning speed model on participants' response time data from the first three training blocks. The interaction between the predictor estimating the size of the response time gain (i.e., second level of the contrast training block) and the predictor variable group provided information concerning whether the response time gain differed between the two groups of participants. The estimate of this interaction was positive but not significant, $\Delta z = 0.21$, $t = 1.44$, 95% profile CI [–0.08, 0.50], $p = .15$; therefore, even if we ignored the statistical problem of the visualization-drivenness of this test, we had no evidence that the response time gain differed between typically developing children and children with DLD.

To further explore the second question, we computed the Pearson correlation coefficient between participants' individual gain in response time and their individual disruption peaks. Because both these individual response time measures included data from Training Block 3, the null hypothesis for the Pearson correlation coefficient was not 0 but .29, that is, $\frac{1}{6}\sqrt{3}$: the correlation between the sum-to-zero contrast of the predictor response time gain ($+\frac{1}{2}$, 0, $-\frac{1}{2}$, 0, 0) and the sum-to-zero contrast of the predictor variable disruption peak (0, 0, $-\frac{1}{3}$, $+\frac{2}{3}$, $-\frac{1}{3}$; see https://osf.io/8a3yv). Thus, we could only conclude that both measures were associated if the confidence interval of the correlation did not include .29. This was the case because the correlation was positive, $r = .67$, 95% CI [.52, .78]. Thus, we could indeed conclude that, on average, children with larger gains in response time from Training Block 1 to Training Block 3 had larger disruption peaks.
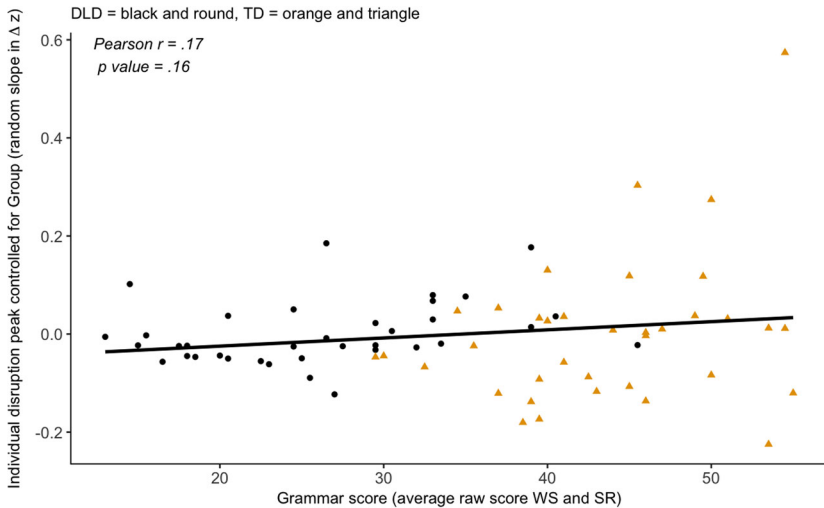
Figure 4 Graphic (descriptive) representation of the relationship between participants' individual disruption peaks and their grammar performance. DLD = developmental language disorder; TD = typically developing; WS = word structure; SR = sentence recall. [Color figure can be viewed at wileyonlinelibrary.com]

Further Exploration of the Link Between Online Statistical Learning and Grammar Performance

For this exploratory analysis, we computed Pearson correlation coefficients between participants' statistical learning performance (individual disruption peaks) and their composite grammar performance score (see Figure 4 for a descriptive visualization of the relationship). We decided to average participants' scores on the sentence recall task and the word structure task because their scores on these tasks were positively correlated, $r(70) = .73$, 95% CI [.65, .82]. Because the individual disruption peaks were extracted from the confirmatory disruption peak model, the individual measure of statistical learning controlled for all predictors that we included in this model (e.g., group, experiment version, verbal working memory, verbal short-term memory). Thus, because the individual measure already controlled for group differences, we estimated the association between NAD learning and grammar for the pooled group of participants rather than for the two participant groups separately. We observed that the correlation between statistical learning and grammar was positive and weak, $r = .17$, 95% CI [–.07, .38]. Thus, we could not conclude that NAD learning, measured through a disruption in response times (and controlled, among other variables, for group status, verbal working memory, verbal short-term
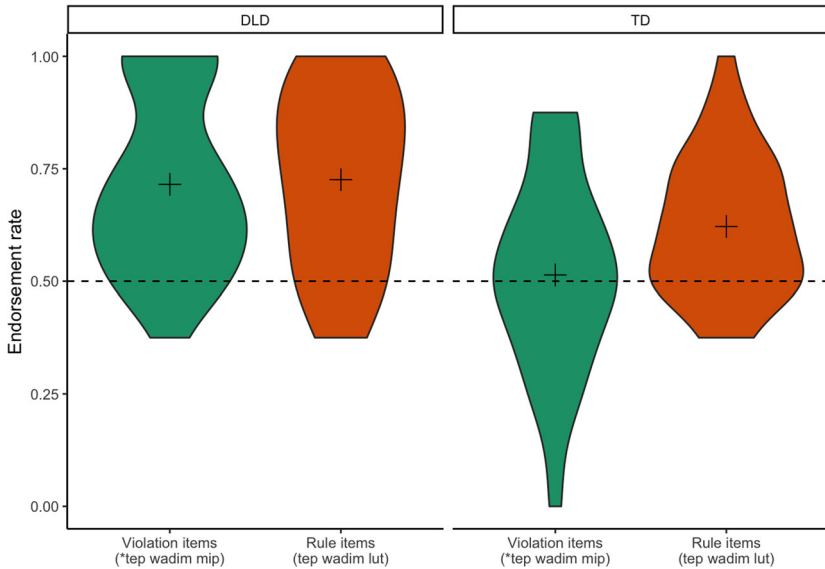
Figure 5 Graphic (descriptive) representation of endorsement rates for item types by group. DLD = developmental language disorder; TD = typically developing. [Color figure can be viewed at wileyonlinelibrary.com]

memory), was associated in our children with expressive morphosyntax, measured through the sentence recall and word structure tasks.

Exploration of the Offline Measure

In a first step, we assessed whether participants endorsed items that were in accordance with the NADs (rule items) more than they endorsed items that violated the NADs (violation items), and referred to this as the rule effect. The model estimated that participants endorsed rule items 1.6 times more often than violation items, but this odds ratio (OR) was not significantly different from 1, log odds = 0.49, z = 1.56, 95% Wald CI for OR [0.9, 3.0], p = .12 (see Table 6 and Figure 5). Therefore, we had no evidence that our offline measure captured children's sensitivity to the NADs. The model estimate for the Rule × Group interaction showed that the rule effect was 1.8 times larger in typically developing children than in children with DLD, but this OR ratio between both groups was not statistically different from 1, log odds = 0.60, z = 1.34, 95% Wald CI for OR ratio [0.8, 4.4], p = .18 (see Table 6). Therefore, we could not conclude that the rule effect differed between children with DLD and typically developing children.

Table 6 Outcome of the generalized linear mixed-effects model for endorsement rate (1,152 observations)

Random effects of subjects ($N = 72$)	SD (log-odds)	Correlation Intercept	Rule	Generalization
Intercept	0.58			
Rule	0.54	–0.42		
Generalization	1.02	–0.27	0.60	
Rule × Generalization	0.69	0.86	–0.26	0.23

Random effects of X elements ($N = 16$)	SD (log-odds)	Correlation Intercept	Group	Experiment version
Intercept	0.44			
Group	0.23	0.17		
Experiment version	0.91	–0.20	0.75	
Group × Experiment Version	0.56	–0.18	–0.68	–0.91

Fixed effect	B_{model} (log-odds)	$B_{transformed}$ (odds)	95% CI (odds)	z	p
Intercept (yes bias)[a]	0.79	2.2	[1.5, 3.1]	4.48	< .001
Group[a]	–0.64	0.5	[0.3, 0.9]	–2.30	.02
Rule[a]	0.49	1.6	[0.9, 3.0]	1.56	.12
Rule × Group[a]	0.60	1.8	[0.8, 4.4]	1.34	.18
Generalization[a]	0.75	2.1	[1.1, 4.1]	2.18	.03
Generalization × Group[a]	0.55	1.7	[0.6, 4.9]	1.04	.30

Note. The log-odds model outputs were transformed to odds, odds ratios (ORs), and OR ratios. The full model outcome (including all predictors) can be found in the R markdown script at https://osf.io/8a3yv and an operationalization of the predictors in Appendix S1. [a]Relevance is exploratory.

One of our criticisms of the use of offline grammaticality judgments has been that children often show a yes bias, as we mentioned previously. And indeed, our model estimated that participants endorsed items (i.e., said "yes I've heard this before") 69% of the time (intercept log odds: 0.79). This is more than one would expect on the basis of chance (50%) and 2.2 times more than the rate of participants' rejection of items, so we could conclude that children showed a yes bias on the offline task, $z = 4.48$, 95% Wald CI probability [61%, 76%], $p < .001$ (see Table 6). The model also estimated that the yes bias was

0.5 times larger (thus 2 times smaller) in typically developing children than in children with DLD, $z = -2.30$, 95% Wald CI for OR [0.3, 0.9], $p = .02$ (see Table 6).

Finally, the model estimated that children endorsed items with familiar X elements 2.1 times more often than items with novel X items, $z = 2.18$, 95% Wald CI for OR [1.1, 4.1], $p = .03$ (see Table 6). The model also estimated that this familiarity effect was 1.8 times larger for typically developing children than for children with DLD, but this difference was not statistically different from 1, $z = 1.04$, 95% Wald CI for OR ratio [0.6, 4.9], $p = .30$ (see Table 6). Our task instructions might have caused this familiarity effect, however (see below).

Discussion
A Small Auditory Statistical Learning Deficit in Children With Developmental Language Disorder

The present study provided new evidence for a statistical learning deficit concerning children's sensitivity to NADs for children with DLD compared to typically developing children. In an artificial language learning experiment, we found that when a long stretch of stimuli with NADs was interrupted by stimuli without dependencies, participants with DLD responded to this interruption with lower disruption peaks than typically developing participants, or that they had no disruption peaks, indicating that children with DLD have an auditory verbal statistical learning deficit. However, the confidence interval of the standardized effect size for this between-group difference ranged from 0.02 to 0.43. These values can be interpreted as a Cohen's d effect size, so that the lower bound of 0.02 standard deviations can be called very small and the upper bound of 0.43 standard deviations as small to medium (Cohen, 1988).

To see how this result fits within the existing literature on statistical learning in children with and without DLD, we have compared the point estimate of our effect size for the between-group difference, which was 0.23 (0.19/0.84), with the range of effect sizes observed in three recent meta-analyses. The meta-analyses differed in whether they examined statistical learning in the visuomotor domain (Lum et al., 2012), the auditory domain (Lammertink et al., 2017), or a combined sample of studies across both domains (Obeid et al., 2016). Also, they differed in whether the studies included in the analyses assessed learning with an online measure such as disruption in response times (Lum et al. 2012), mostly offline measures (Lammertink et al., 2017), or a mixture of online and offline measures (Obeid et al., 2016). In sum, we observed that (a) our point estimate of 0.23 fell within the limits of the confidence interval for (and was thus

compatible with) the statistical learning deficit—which ranged from 0.072 to 0.584—reported in Lum et al. (across eight studies); (b) our point estimate was smaller than the lower bound of the confidence interval reported in Lammertink et al. (0.36 across 10 studies); and (c) our point estimate was also smaller than the lower bound of the confidence interval reported in Obeid et al. (0.276 across 14 studies). From this, we speculate that it is rather the method of measuring statistical learning (online vs. offline) than the domain in which learning takes place (visuomotor vs. auditory) that impacts the size of the reported deficit.

Offline grammaticality judgments (as commonly used in the word segmentation and artificial grammar studies that were included in the meta-analyses by Lammertink et al., 2017, and Obeid et al., 2016) apparently lead to a larger difference between children with and without DLD than online measures of learning. Other than the modality and/or method of measuring statistical learning, the type of statistical structure to be learned (e.g., adjacent, nonadjacent, hierarchical) may also affect the size of the statistical learning deficit. Given that the detection of NADs is thought to be more cognitively demanding than the detection of adjacent dependencies (Wilson et al., 2018), the size of the NAD learning deficit observed in the present study may be surprisingly small (i.e., this would suggest an adjacent dependency learning deficit to be even smaller). We speculate, however, that learning the NADs was relatively easy for both groups of participants because we optimized the NAD learning conditions in the present experiment (see Wilson et al., 2018, for an overview on the constraints of NAD learning). That is, (a) we decreased the transitional probability between adjacent elements (thereby increasing the saliency of NADs) by using 24 different X elements; (b) we made the NAD elements (*tep* and *lut*; *sot* and *mip*) perceptually more similar to each other than to the intervening X elements; and (c) the NAD elements were positioned at the start and end of the sequence making them easier to detect (referred to as edge effects in Wilson et al., 2018). Because we cannot make a direct comparison between the size of the NAD learning deficit (present study) and the size of an adjacent dependency learning deficit (estimate not available; the meta-analyses cited above contained studies with a mixture of dependency types), in future studies researchers may want to use within-subject designs to further investigate how the type of statistical structure relates to the size of the statistical learning deficit in children with DLD.

Measuring Nonadjacent Dependency Learning in Children
The use of online measures of statistical learning in the auditory domain is relatively new. Therefore, new measures keep emerging. For example, in a

recently published paper Kuppuraj et al. (2018) showed that adults' sensitivity to sequences, including NADs, in the auditory domain can also be assessed through a difference in slopes at the transition point between sequenced and nonsequenced items. A slope difference may be expected if participants exhibit statistical learning (large negative slope) during the pre-disruption blocks, and participants do not exhibit it (0 or perhaps slightly negative slope) during the disruption block. By contrast, a difference in disruption peak height (as used in the present study) may be expected if participants are better at predicting regularities during sequenced blocks than during the disruption block. Both effects are likely to play a role, and our exploratory results suggest that the effects are associated, but their relative strengths determine which of the two will be easier to detect in an experiment. Determining under what circumstances which method of measuring fits best with the existing literature on the online measurement of statistical learning (e.g., via Monte Carlo simulations) is beyond the scope of the present article but may be relevant for future work.

Given that our online measure of NAD learning was relatively new, it may be good to address the reliability and validity of the measure. We derived indications of the reliability from different sources. First, the widths of the reported confidence interval around the standardized effect size for our confirmatory measure ranged from small to medium, indicating moderate reliability (the smaller the width, the more reliable a measure is). Second, by using a linear mixed-effects model with a random intercept for X element and with random slopes for X element, we could conclude that the reported effects generalize to the population of all possible X elements and thus that the size of the disruption peak was not specific to the X elements in the artificial language used in the present study. Finally, the online NAD measure (disruption peak) had a split-half reliability (Spearman-Brown corrected) of .79, with a 95% confidence interval ranging from .66 to .87 (see our R markdown script at https://osf.io/8a3yv for computation of the split-half reliability). As to the validity of our results, the present study combined two measures that are commonly used to measure the construct of statistical learning. First, disruption peaks have been shown to be a valid measure of people's sensitivity to statistical regularities in serial reaction time studies (e.g., Conway, Arciuli, Lum, & Ullman, 2019; Lum et al., 2012). Second, NAD learning studies have shown that infants and adults learn structure from exposure to miniature artificial languages comparable to the language used in the present study (Gómez, 2002). Finally, Lammertink, van Witteloostuijn et al. (2018) showed that the combination of the measures from the design as used in the present study led to a valid measure of NAD learning in primary-school-aged children.

Alternative Explanations

Rather than a statistical learning deficit, an alternative explanation for the difference observed between children with and without DLD in auditory statistical learning studies may be that limitations in verbal short-term memory, verbal working memory, or processing speed in children with DLD hinder their detection of NADs. However, our statistical analysis detected a difference between children with and without DLD even when we controlled for verbal short-term memory and for verbal working memory. Therefore, we argue that reduced memory capacity is not the limiting factor in children's detection of NADs. Furthermore, visual inspection of the participating children's raw response times (in milliseconds) to the target and nontarget items in the first training block (Figure 2) may suggest that participants with DLD and typically developing participants responded equally fast in this first block. If participants with DLD had required more time for processing the auditory stimuli, then one would have already expected to observe slower response times in this first training block. Thus, from this observation, we also speculate that differences in processing time are not the limiting factor in children's detection of NADs. Finally, we found no evidence that participants with DLD made more errors during the online phase of the experiment, which means that we have no indirect evidence that children with DLD had more difficulties with the task.

Because we found that NAD learning differed based on general language proficiency at the group level (DLD vs. typically developing), we further explored if sensitivity to NADs was correlated with participants' knowledge of morphological and morphosyntactic rules at the individual level. We found no evidence for (or against) such a relationship. Of course, the sentence recall task and the word structure task with which we assessed participants' morphosyntactic and morphological knowledge are not pure measures of children's sensitivity to NADs in natural language. For example, there is some debate about whether the sentence recall task taps solely into morphosyntactic ability or whether task results also depend on other cognitive processes such as working memory (Frizelle, O'Neill, & Bishop, 2017). As for the word structure task, this task assesses children's knowledge of relatively simple items that are highly frequent in Dutch (the task has been developed for children between 5 and 8 years of age). Therefore, it could well be the case that children retrieve the correct forms of the items from their declarative memory instead of using morphological rules. This may mean that the word structure task is more sensitive to rote learning strategies rather than to statistical or rule learning strategies.

The number of participants tested is typically small in clinical studies. Consequently, the power of clinical studies may be too low to detect the effects

under examination. However, we have two reasons to believe that the present study was sufficiently powered to detect the effects under examination. First, in comparison to serial reaction time task studies, the number of participants with DLD whom we tested for the present study was relatively large. In the serial reaction time task studies (approximately 11 studies in total), the number of participants with DLD has ranged from 14 to 48, with only two studies reporting more than 36 participants (Conti-Ramsden, Ullman, & Lum, 2015; Hsu & Bishop, 2014a). Second, we did detect an effect in our online measure. This indicates that we tested a sufficient number of participants to detect a difference in NAD learning between participants with and without DLD. Also, the confidence interval for this effect had a small range. In underpowered studies, this range would be large.

A limitation of the present study is that our offline forced-choice task measure could not detect NAD learning. Instead of asking participants whether they thought the utterance with which we presented them followed the rules of the language, we asked them whether they had heard the utterance before. This formulation may have changed the nature of the offline task, making it a recognition task rather than a grammaticality judgment task. As such, it may be no surprise that participants showed a familiarity preference (i.e., they were more likely to respond yes to items with familiar X elements than items with novel X elements). Given this limitation, we deem it impossible to draw any conclusions from our offline measure of learning.

Conclusion

We would like to end our discussion with some words about why the study of NAD learning in children with DLD is relevant for professionals and researchers working with these children. Our discussion of these clinical implications is, of course, speculative. Before any firm conclusions can be drawn about the clinical relevance of the potentially small NAD learning deficit in children with DLD, future studies may first want to further develop the measure of NAD learning. Nevertheless, if the small magnitude of the auditory NAD learning deficit in DLD is replicated, then one may argue that it may be more effective to focus on the improvement of other skills important for children's language development (e.g., phonological processing, phonological working memory) rather than to focus on the development of therapies that aim to improve children's statistical learning ability. For example, a meta-analysis by Graf Estes, Evans, and Else-Quest (2007) showed that children with DLD performed on average 1.27 standard deviations (95% CI [1.15, 1.39]) below their typically developing peers on a nonword repetition task. This effect size

was larger than the effect size observed in the present study and also larger than the effect sizes reported by Lammertink et al. (2017), Lum et al. (2014), and Obeid et al. (2016) in their meta-analyses of statistical learning in children with DLD. Thus, the gains in children's language ability may be higher for therapies that focus on children's phonological skills than for therapies that focus on their detection of statistical regularities.

Alternatively, because the auditory verbal statistical learning deficit in children with DLD is small, the deficit could potentially be easily resolved if ways are found to facilitate the detection of NADs in children with DLD at an early age. Recently, Plante and Gómez (2018) made a similar argument and provided concrete examples for incorporating the principles of statistical learning in already existing language interventions for children with DLD. For example, it has been suggested that variability in the nontarget structure (i.e., the X elements in NAD pairs) facilitates the detection of regularities in the input (Gómez, 2002; Plante et al., 2014). Such findings are encouraging, but also assume (and require) that children with DLD apply a statistical rather than a rote learning strategy in a natural (rather than artificial) language learning context. Hsu and Bishop (2014b), for example, concluded that using a statistical learning strategy may be problematic for children. They observed that, in a natural language context, children tend to rely more on a rote learning strategy. Therefore, the first step may be to investigate how educators can encourage children with DLD to rely on statistical cues in their native language input before they incorporate the principles of statistical learning into the existing language interventions. In conclusion, although the present study provided new evidence for a statistical learning deficit specific to NADs in children with DLD compared to the statistical learning in typically developing children, we acknowledge that this deficit is probably small in size.

Final revised version accepted 11 June 2019

Notes

1 The present study was part of a larger research project on the relationship between statistical learning, grammar, and literacy acquisition in children. Consequently, we have also reported data from the same group of participants with DLD and typically developing participants in Lammertink, Boersma et al. (2018) and Lammertink et al. (2019b). Van Witteloostuijn, Boersma, Wijnen, and Rispens (accepted, 2019) have also described a subset of the typically developing participants in separate studies, with different research questions, and a different clinical group (developmental dyslexia).

2 In a first step, we fitted an online disruption peak model that included a per-subject random slope for the interaction between the variables block and targetness and a per-item slope for the interaction between the variables experiment version and group status as well. However, the profile method failed to compute a confidence interval for our predictor of interest for this maximal model. When we removed the near-to-perfect correlation between the interactions in our random effects structure (Bates et al., 2018), the profile method worked. We were allowed to remove these interactions because they were not of interest to our confirmatory research question (e.g., we report no *p* values for them). For more details, see the R markdown file in the Supplementary Information online.

References

Ambridge, B., & Lieven, E. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge, UK: Cambridge University Press.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics from the behavioral sciences*. New York, NY: Palgrave Macmillan.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. (2018). *Parsimonious mixed models*. https://arxiv.org/pdf/1506.04967.pdf

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

Bialystok, E. (1986). Factors in the growth of linguistic awareness. *Child Development*, *57*, 498–510. https://doi.org/10.2307/1130604

Bishop, D. (2003). *Test for reception of grammar* [Measurement instrument]. London, UK: Pearson.

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*, 1068–1080. https://doi.org/10.1111/jcpp.12721

Braams, T., & de Vos, T. (2015). *Schoolvaardigheidstoets Spelling* [Dutch Spelling test: Measurement instrument]. Amsterdam, Netherlands: Boom test uitgevers.

Brus, B., & Voeten, M. (1979). *Een-Minuut-Test* [One minute test: Measurement instrument]. Amsterdam, Netherlands: Pearson.

Chevrie-Muller, C., Maillart, C., Simon, A., & Fournier, S. (2010). *Batterie langage oral, langage écrit, mémoire, attention* (2ème éd.) [Oral language, written language, memory, attention test battery: Measurement instrument]. Paris, France: Édition du centre de psychologie appliquée.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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*, 1–11. https://doi.org/10.3389/fpsyg.2015.01090

Conway, C., Arciuli, J., Lum, J., & Ullman, M. (2019). Seeing problems that may not exist: A reply to West et al.'s (2018) questioning of the procedural deficit hypothesis. *Developmental Science*. Published online 11 February 2019. https://doi.org/10.1111/desc.12814

Desmottes, L., Meulemans, T., & Maillart, C. (2016). 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

Duinmeijer, I. (2016). *Persistent grammatical difficulties in specific language impairment: Deficits in knowledge or in knowledge implementation*? (Unpublished doctoral dissertation). University of Amsterdam, Amsterdam, Netherlands.

Ebert, K., & Kohnert, K. (2011). Sustained attention in children with primary language impairment: A meta-analysis. *Journal of Speech, Language and Hearing Research*, *54*, 1372–1384. https://doi.org/10.1044/1092-4388(2011/10-0231)

E-prime (Version 2.0) [Computer Software]. (2012). Pittsburgh, PA: Psychology Software Tools.

Evans, J., Saffran, J., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*, 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*, 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*, 1435–1457. https://doi.org/10.1017/S0305000916000635

Gabriel, A., Maillart, C., Guillaume, M., Stefaniak, N., & Meulemans, T. (2011). Exploration of serial structure procedural learning in children with language impairment. *Journal of the International Neuropsychological Society*, *17*, 336–343. https://doi.org/10.1017/S1355617710001724

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

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

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436. https://doi.org/10.1111/1467-9280.00476

Graf Estes, K., Evans, J., Else-Quest, N. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *50*, 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Grunow, H., Spaulding, T., Gómez, R., & Plante, E. (2006). The effects of variation on learning word order rules by adults with and without language-based learning disabilities. *Journal of Communication Disorders*, *39*, 158–170. https://doi.org/10.1016/j.jcomdis.2005.11.004

Haebig, E., Saffran, J., & Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *The Journal of Child Psychology and Psychiatry*, *58*, 1251–1263. https://doi.org/10.1111/jcpp.12734

Hamrick, P., Lum, J., & Ullman, M. (2017). 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

Hsu, H., & Bishop, D. (2011). Grammatical difficulties in children with specific language impairment: Is learning deficient? *Human Development*, *53*, 264–277. https://doi.org/10.1159/000321289

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

Hsu, H. J., & Bishop, D. V. M. (2014b). Training understanding of reversible sentences: A study comparing language-impaired children with age-matched and grammar matched controls. *PeerJ*, *3*, 1–23. https://doi.org/10.7717/peerj.656

Hsu, H., Tomblin, J., & Christiansen, M. (2014). Impaired statistical learning of non-adjacent dependencies in adolescents with specific language impairment. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00175

Iao, L.-S., Ng, L., Wong, A., & Lee, O. (2017). Nonadjacent dependency learning in Cantonese speaking children with and without specific language impairment. *Journal of Speech, Hearing and Language Research*, *60*, 694–700. https://doi.org/10.1044/2016_JSLHR-L-150232

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.

Khomsi, A. (2001). *Évaluation du langage oral* [Oral language evaluation: Measurement instrument]. Paris, France: Édition du centre de psychologie appliquée.

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*, 794–829. https://doi.org/10.1080/01690965.2010.493735

Krok, W., & Leonard, L. (2015). Past tense production in children with and without specific language impairment across Germanic languages: A meta-analysis. *Journal of Speech, Language and Hearing Research*, *58*, 1326–1340. https://doi.org/10.1044/2015_JSLHR-L-14-0348

Kuppuraj, S., Duta, M., Thompson, P., & Bishop, D. (2018). Online incidental statistical learning of audiovisual word sequences in adults: A registered report. *Royal Society Open Science*, *5*, 171678. https://doi.org/10.1098/rsos.171678

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*, 3474–3486, https://doi.org/10.1044/2017_JSLHR-L-16-0439.

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2018, June). *Individual differences in children with developmental language disorder: Associations between visual statistical learning, reading and phonological processing*. Paper presented at the Child Language Symposium, Reading, UK. https://doi.org/10.1017/S0142716418000577

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019a, June). *The (missing) link between statistical learning and grammar knowledge.* Poster presented at the Interdisciplinary Advances in Statistical Learning Conference, San Sebastian, Spain. https://doi.org/10.13140/RG.2.2.23251.53285

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019b, June). *A statistical learning deficit in children with developmental language disorder: Investigating the role of modality and type of dependency*. Paper presented at the Interdisciplinary Advances in Statistical Learning Conference, San Sebastian, Spain.

Lammertink, I., van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2018). Auditory statistical learning in children: Novel insights from an online measure. *Applied Psycholinguistics*, *40*, 279–302.

Lecocq, P. (1998). *Épreuve de compréhension syntaxico-sémantique: Adaptation française du TROG: Reception of Grammar Test* [Test of syntax-semantic comprehension: French adaptation of the TROG: Reception of Grammar Test; Measurement instrument]. Villeneuve d'Ascq, France: Presses universitaires du Septentrion.

López-Barroso, D., Cucurell, D., Rodríguez-Fornells, A., & de Diego-Balaguer, R. (2016). Attentional effects on rule extraction and consolidation from speech. *Cognition*, *152*, 61–69. https://doi.org/10.1016/j.cognition.2016.03.016

Lukács, Á., & Kemény, F. (2014). Domain-general sequence learning deficit in specific language impairment. *Neuropsychology*, *28*, 472–483. https://doi.org/10.1037/neu0000052

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*, 1–10. https://doi.org/10.1016%2Fj.cortex.2013.10.011

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

Manly, T., Robertson, I., Anderson, V., & Nimmo-Smith, I. (2010). *Test of everyday attention for children: Manual, Dutch version* [Measurement instrument]. Amsterdam, The Netherlands: Pearson.

Mayor-Dubois, C., Zesiger, P., Van der Linden, M., & Roulet-Perez, E. (2014). Nondeclarative learning in children with specific language impairment: Predicting regularities in the visuomotor, phonological, and cognitive domains. *Child Neuropsychology*, *20*, 1–9. https://doi.org/10.1080/09297049.2012.734293

Misyak, J., & Christiansen, M. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 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*, 1–9. https://doi.org/10.3389/fpsyg.2010.00031

Montgomery, J., Evans, J., & Gillam, R. (2018). Memory and language in children with SLI. In T. Alloway (Ed.), *Working memory and clinical developmental disorders* (pp. 22–37). London, UK: Routledge.

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

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

Plante, E., & Gómez, R. (2018). Learning without trying: The clinical relevance of statistical learning. *Language, Speech, and Hearing Services in Schools*, *49*, 710–722. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0131

Plante, E., Ogilvie, T., Vance, R., Aguilar, J., Dailey, N., Meyers, C., . . . Burton, R. (2014). Variability in the language input to children enhances learning in a treatment context. *American Journal of Speech-Language Pathology*, *23*, 530–545. https://doi.org.10.1044/2014_AJSLP-13-0038.

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: 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*. San Antonio, TX: Harcourt.

Renfrew, C. (2003). *The action picture test* (4th ed.) [Measurement instrument]. Oxford, UK: Speechmark.

Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL* [Measurement instrument]. Amsterdam, Netherlands: Harcourt.

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

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

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Sociaal en Cultureel Planbureau (2017, February). *Statusscores 2016* [report Social and Cultural planning]. http://www.scp.nl/Formulieren/Statusscores_opvragen

Spit, S., & Rispens, J. (2018). On the relation between procedural learning and syntactic proficiency in gifted children. *Journal of Psycholinguistic Research*, *48*, 417–429. https://doi.org/10.1007/s10936-018-9611-6

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

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

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019, accepted). Statistical learning abilities of children with developmental dyslexia across three experimental paradigms. *PLOS ONE*.

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019, June). *The contribution of individual differences in statistical learning to reading and spelling performance in children with and without dyslexia*. Poster presented at the Interdisciplinary Approaches to Statistical Learning (IASL), San Sebastian, Spain. https://doi.org/10.13140/RG.2.2.24037.96483

Vuong, L., Meyer, A., & Christiansen, M. (2015). Concurrent statistical learning of adjacent and nonadjacent dependencies. *Language Learning*, *66*, 8–30. https://doi.org/10.1111/lang.12137

Wijnen, F. (2013). Acquisition of linguistic categories: Cross-domain convergences. In J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech and language: Exploring the evolution of mind and brain* (pp. 157–177). Cambridge, MA: 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*. Advanced online publication. https://doi.org/10.1111/tops.12381

Supporting Information

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

Appendix S1. Operationalization of Contrasts.

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

Children With Developmental Language Disorder Have Difficulties With Picking Up Language "Rules" From Exposure to Language

What This Research Was About and Why It Is Important

Developmental language disorder (DLD) manifests itself (among other ways) as difficulties that children have learning grammar rules in their native language. Children with DLD experience problems with their social interaction and delays in education progress. The prevalence of DLD is estimated at 7%, which means that there is approximately one child with DLD in every classroom. Because the difficulties in language learning that these children experience have no clear cause such as low intelligence, brain damage, or hearing impairment, it is important to understand potential other causes of DLD so that the impact of DLD on children's language development might be mitigated. Previous research has shown that children are generally sensitive to regularities in their language input, which helps them learn their native language. For example, in English, the pronoun *he* frequently co-occurs with [verb]–*s* in the present tense, as in *he* walk*s*, *he* talk*s*, and *he* eat*s*. Often without conscious awareness, children detect and keep track of these co-occurrences, which guides them in learning the systems (patterns or "rules") underlying English grammar. And this ability—that is, being sensitive to linguistic regularities in a language—has been proposed as one possible difference between children with DLD and typically developing children. Therefore, in this study, the researchers aimed to understand whether children with DLD are as sensitive as their typically developing peers to linguistic regularities. The researchers found that the children

with DLD indeed had more difficulty keeping track of linguistic regularities compared to typically developing children, suggesting that this ability may be one of the reasons why children with DLD have difficulties learning grammar systems.

What the Researchers Did
- The researchers exposed 36 children with DLD and 36 children without DLD (i.e., typically developing children) to a novel nonexisting language. All children were native speakers of Dutch, between 8 and 12 years old.
- Each utterance in the nonexisting language consisted of three words. Unbeknown to the children, the first word of the utterance (which was either *tep* or *sot*) "predicted" the third word (which was either *lut* or *mip*, respectively), as in *tep wadim lut* and *sot kasi mip*. In other words, "*tep* and *lut*" as well as "*sot* and *mip*" always went together in an utterance.
- The children heard these utterances, presented to them through a mini cartoon on a tablet computer, and were then asked to press a green or red button. The color of the button they had to press depended on the third word of the utterance (e.g., they had to press the green button upon hearing *lut*). The children's response speed in detecting the third word was indicative of whether they had learned the co-occurrence of the words (*tep . . . lut* and *sot . . . mip*). If they had learned that the first word predicted the third word, they would be quicker to press the correct button about the third word.

What the Researchers Found
- The response pattern of the children with DLD differed from those of the typically developing children, suggesting that the children with DLD were less sensitive to the word co-occurrences than the typically developing children.
- However, this difference between the two child groups was small in magnitude.

Things to Consider
- In future, it might be worth exploring ways to assist children with DLD with detecting linguistic regularities in the language and using these regularities for learning grammar systems.
- However, because the differences between children's response patterns were small, researchers might instead consider targeting other language skills that are impaired in children with DLD, such as comprehension of speech, in order to help them acquire their native language more efficiently.

- Taken together, the results point to one of potentially many differences underlying language learning by children with DLD and by typically developing peers.

How to cite this summary: Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Children with developmental language disorder have difficulties with picking up language "rules" from exposure to language. *OASIS Summary* of Lammertink et al. in *Language Learning*. https://oasis-database.org

This summary has a CC BY-NC-SA license.