Applied Psycholinguistics **40** (2019), 279–302
doi:10.1017/S0142716418000577

Auditory statistical learning in children: Novel insights from an online measure

IMME LAMMERTINK, MEREL VAN WITTELOOSTUIJN, and PAUL BOERSMA
Amsterdam Center for Language and Communication, University of Amsterdam

FRANK WIJNEN
Utrecht Institute of Linguistics OTS, Utrecht University

JUDITH RISPENS
Amsterdam Center for Language and Communication, University of Amsterdam

Received: December 28, 2017 Revised: June 5, 2018 Accepted: August 22, 2018

ADDRESS FOR CORRESPONDENCE
Imme Lammertink, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam.
E-mail: i.l.lammertink@uva.nl

ABSTRACT
Nonadjacent dependency learning is thought to be a fundamental skill for syntax acquisition and often assessed via an offline grammaticality judgment measure. Asking judgments of children is problematic, and an offline task is suboptimal as it reflects only the outcome of the learning process, disregarding information on the learning trajectory. Therefore, and following up on recent methodological advancements in the online measurement of nonadjacent dependency learning in adults, the current study investigates if the recording of response times can be used to establish nonadjacent dependency learning in children. Forty-six children (mean age: 7.3 years) participated in a child-friendly adaptation of a nonadjacent dependency learning experiment (López-Barroso, Cucurell, Rodríguez-Fornells, & de Diego-Balaguer, 2016). They were exposed to an artificial language containing items with and without nonadjacent dependencies while their response times (online measure) were measured. After exposure, grammaticality judgments (offline measure) were collected. The results show that children are sensitive to nonadjacent dependencies, when using the online measure (the results of our offline measure did not provide evidence of learning). We therefore conclude that future studies can use online response time measures (perhaps in addition to the offline grammaticality judgments) to further investigate nonadjacent dependency learning in children.

Keywords: grammaticality judgment; language development; nonadjacent dependency learning; response time; statistical learning

Statistical learning, the ability to detect structure in the environment, plays a key role in the development of language, perception, motor skills, and social behavior (cf. Perruchet & Pacton, 2006). It is not surprising, then, that an increasing number of studies investigate the relation between individual statistical learning

performance and cognitive development. A particular type of statistical learning is nonadjacent dependency learning (NAD-learning). Nonadjacent dependencies are amply present in natural language. Consider, for example, the relation between the functional elements *is* and *ing* across interleaved lexical elements in *Grandma* ***is*** *sing****ing*** (example taken from Sandoval & Gómez, 2013). For this reason, NAD-learning is thought to be fundamental for syntax acquisition (see review by Erickson & Thiessen, 2015), and in adults, sensivity to nonadjacent dependencies has shown to predict online processing of long-distance dependencies in relative clauses (Misyak, Chirstiansen, & Tomblin, 2010).

However, the generally used measure of NAD-learning, an offline group-level grammaticality judgment score (Gómez, 2002), is problematic when evaluating the learning ability of individuals as this offline measure only quantifies the extent of learning after a specific period of time (i.e., *what* is learned). It does not provide insight in the speed of learning, nor can it disentangle statistical learning from other processes potentially impacting the offline measure, such as encoding, memory capacity, and decision-making biases (i.e., *how* learning occurs; Siegelman, Bogaerts, Kronenfeld, & Frost, 2017). Therefore, a growing body of research stresses the importance of using measures that provide information on the individual learning trajectory and/or the various processes involved in NAD-learning (López-Barroso, Cucurell, Rodríguez-Fornells, & de Diego-Balaguer, 2016; Misyak et al., 2010).

In the classical offline NAD-learning task, participants are exposed to strings of an artificial language. The strings consist of three pseudowords that, unbeknownst to the participant, contain nonadjacent dependencies. The strings have the form *a* X *b, c* X *d, e* X *f* with the initial and final elements forming a dependency pair. The intervening X-elements vary and are usually taken from a pool of different pseudowords (e.g., *wadim, kasi*; Gómez, 2002). After a certain period of exposure to the artificial language, participants perform a grammaticality judgment task in which they are tested with strings that either conform to the nonadjacent dependency rules or violate the nonadjacent dependency rules. If participants' proportion of correct answers on the grammaticality judgment task exceeds chance level, it is concluded that they are sensitive to the nonadjacent dependency rules. As we will argue later, this reliance on the offline measure *only* is problematic as it might not fully reflect participants' (unconscious) acquired knowledge of the nonadjacent dependencies. It also disregards all information regarding the learning dynamics during exposure to the novel language.

As an increasing number of researchers stresses the importance of measuring statistical learning in a different way than by grammaticality judgments, several different measures have been proposed (e.g., the statistically induced chunking recall task; see Isbilen, McCauley, Kidd, & Christiansen, 2017). In the current paper we focus on the collection of response times (RTs) as an online measure of NAD-learning. The use of RTs as an online measure of learning has its roots in the serial reaction time (SRT) literature (Nissen & Bullemer, 1987). In the SRT task, RTs have been shown to successfully track participants' (both adults and primary school-age children; Thomas & Nelson, 2001) online learning of visuomotor sequences. In the original version of the task, participants have to respond to a visual stimulus appearing in one of four locations on a screen.

Participants' RTs in sequenced blocks (stimuli follow a fixed sequence) are compared to their RTs in random nonsequenced blocks (stimuli appear in random order). The typical result is that participants respond faster in sequenced blocks than in random blocks, and this effect is taken as evidence for implicit learning of the sequence. Following this pattern of results, two recent studies transformed the SRT task into an online NAD-learning experiment. Both studies successfully showed that RTs can be used to track NAD-learning in the auditory domain in adults (López-Barroso et al., 2016; Misyak et al., 2010). Of these two studies, the latter resembles the SRT paradigm most closely. Misyak et al. designed a cross-modal paradigm in which participants were auditorily exposed to strings consisting of three pseudowords and three dependency pairs (*a* X *b*, *c* X *d*, *e* X *f*; Gómez, 2002). Participants were simultaneously presented with six printed pseudowords on a screen and asked to click as fast as possible on the nonsense word that matched the auditorily presented word. Thus, for example, participants heard the string *pel wadim rud*, then the participant first clicked *pel* upon hearing *pel*, then *wadim* upon hearing *wadim*, and finally *rud* upon hearing *rud.* Similarly, as in the SRT paradigm, the sequenced blocks (i.e., blocks containing nonadjacent dependencies) were temporarily disrupted by one nonsequenced block in which the strings violated the nonadjacent dependency rules (e.g., **a* X *d*,**a* X *f*). Misyak et al. showed that participants' RTs were slower in the nonsequenced block than in the surrounding sequenced blocks, confirming that adults are sensitive to the nonadjacent dependency pairs. Whereas this cross-modal design works well with adults, it is difficult to use with (young) children, as well as with participants from language-impaired populations as the task requires good reading skills. Another auditory online NAD-learning task, developed by López-Barroso et al. (2016), remedies this shortcoming.

López-Barroso et al. designed a NAD-learning experiment in which the SRT task is combined with a word-monitoring task (for a comparable design in another type of auditory statistical learning task, see Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015). As in Misyak et al. (2010) and in the classical NAD-learning studies (Gómez, 2002), adults were exposed to artificial language strings that were generated according to nonadjacent dependency rules. Adults had to press a green or red button upon hearing a specific target item, rendering the task completely auditory. The targets were always the final elements of the nonadjacent dependency pairs (*a* X *b*, *c* X *d*). After a certain amount of exposure to the rule items (sequenced blocks), adults were presented with strings in which the NAD rules were disrupted. For example, items contained the *b*-element as the final element, but this was not preceded by the *a*-element as before, and so these items are analogous to the random block in an SRT task. In analogy with the SRT task, adults' RTs to target elements were shorter in the nonadjacent dependency items compared to the random items, reflecting anticipatory word monitoring, and the authors therefore conclude that RTs can be used to track adults' sensitivity to nonadjacent depedencies.

To the best of our knowledge, no published studies have tracked auditory NAD-learning online in primary school-aged children, and only one published study reports on offline NAD-learning in primary school-aged children (Iao, Ng, Wong, & Lee, 2017). As the use of online measures of NAD-learning is relatively new,

this lack of online measures in primary school-aged children is not surprising. The low number of studies reporting on offline measures, however, is surprising as there is ample evidence of offline auditory NAD-learning in *infants* (e.g., 4-month-olds: Friederici, Mueller, & Oberecker, 2011; 18-month-olds: Gómez, 2002; 15- and 18-month-olds: Gómez & Maye, 2005) and *adults* (e.g., Gómez, 2002; Newport & Aslin, 2004, Onnis, Monaghan, Christiansen, & Chater, 2004). This could be because the generally used offline measures of NAD-learning (grammaticality judgments) are difficult to administer to children of this particular age. NAD-learning in infants is assessed via the headturn-preference procedure, a procedure unsuitable for older children (Cristia, Seidl, Singh, & Houston, 2016). As for the offline grammaticality judgment score of NAD-learning in adults, some shortcomings were already mentioned above, but compared to adults, the offline grammaticality judgment measures of NAD-learning might be even more problematic in children as such measures involve some form of metalinguistic awareness that children acquire relatively late (Bialystok, 1986) and that requires more than the language representation alone (e.g., attention and executive functioning). In yes/no grammaticality judgment tasks, children often show a yes bias: they simply accept close-enough descriptions or they reject strings for reasons unrelated to the dependency rules (Ambridge & Lieven, 2011). The two-alternative forced-choice design (choosing one option out of two possibilities) forces children to make a selection when they might think that both (or neither) options are correct (McKercher & Jaswal, 2012). For these reasons, the child's offline judgment might not always reflect sensitivity to the nonadjacent dependencies.

THE CURRENT STUDY

Prompted by the absence of online measures of NAD-learning in primary school-aged children and by the low number of offline NAD-learning measures in this age range, our aim was to investigate whether primary school-aged children are sensitive to nonadjacent dependencies in an artificial language. In order to investigate this, two research questions were formulated:

1. Can we measure primary school-aged children's sensitivity to nonadjacent dependencies online by means of recording RTs?
2. Can we measure primary school-aged children's sensitivity to nonadjacent dependencies offline by means of an offline grammaticality judgment task?

Similarly to conventional offline NAD-learning experiments and to the online auditory NAD-learning experiment of López-Barroso et al. (2016), we exposed children to strings of an artificial language that, unbeknownst to the children, were generated according to a rule (i.e., the strings have an *a* X *b* structure in which the *a*-element and the *b*-element always co-occur; see Gómez, 2002). Children performed a word-monitoring task that allowed us to measure children's RTs to the *b*-elements. After a certain amount of exposure to the nonadjacent dependencies, we presented items that were discordant with the nonadjacent dependencies (disruption

block). In analogy with the SRT task, we predict that if children are sensitive to the nonadjacent dependecies, their RTs to the *b*-element should increase in the disruption block relative to the preceding training block and decrease again, after the disruption block, when rule-based items return in the recovery block. After the online measurement of learning, the children took part in an offline measurement of learning (a two-alternative grammaticality judgment task), and then their explicit knowledge of the rules was evaluated by means of a short questionnaire.

Finally, we explored the relationship between the online measure and offline measure of NAD-learning. We hypothesize that if both measures reflect sensitivity to NADs, children's RTs to the target items will increase in the disruption block relative to the surrounding blocks and they will perform above chance level on the grammaticality judgment task. However, as grammaticality judgments are likely problematic for children, it is possible that we would observe a discrepancy between the two measures.

METHOD

Participants

Fifty-four native Dutch-speaking primary school-aged children participated in the experiment. Eight were excluded for a variety of reasons: equipment error ($N = 1$), not finishing the experiment ($N = 3$), or because overall accuracy in the online word monitoring task was lower than 60% ($N = 4$). As a result, 46 children were included in the final analysis (female = 22, male = 24; mean age = 7.3 years; range = 5.9–8.6 years). No hearing, vision, language, or behavioral problems were reported by their teachers. Children were recruited via four different primary schools across the Netherlands. Approval was obtained from the ethics review committee of the University of Amsterdam, Faculty of Humanities.

Apparatus

The experiment was presented on a Microsoft Surface 3 tablet computer using E-prime 2.0 (2012) software (Psychology Software Tools, Pittsburgh, PA). RTs were recorded with an external button box attached to this computer. The auditory stimuli were played to the children over headphones (Senheiser HD 201).

Materials and procedure

The task. The structure of our NAD-learning experiment is similar to that of conventional NAD-learning experiments. Children were exposed to an artificial language that contained two nonadjacent dependency rules (*tep* X *lut* and *sot* X *mip*). This exposure phase was followed by a grammaticality judgment task and a short questionnaire that assessed awareness of the nonadjacent dependencies. In contrast to conventional NAD-learning experiments, however, children performed a word monitoring task, which allowed us to track children's online learning trajectory by means of a RT measure. To this end, we designed a child-friendly adaptation of an online NAD-learning experiment that was administered

to adults (López-Barroso et al., 2016). As in conventional NAD-learning tasks, children were not informed about the presence of any regularities in the artificial language, rendering the task an incidental learning task.

The word monitoring part of the experiment was framed as a game in which children were instructed to help *Appie* (a monkey) on picking bananas. *Appie* taught the children that they would hear utterances consisting of three nonexisting words (pseudowords) and that they had to press the green button, as quickly as possible, when they heard the specific target word and the red button when none of the three words was the specific target word. In addition, *Appie* told the children that it was important to pay attention to all three words in the utterances, because questions about the utterances would follow at the end (i.e., the grammaticality judgment task). Children were told only that questions would follow, but they were not informed on the nature of the questions. Two versions of the experiment were created, with either *lut* (Version 1) or *mip* (Version 2) as the target word. The target word remained the same across the whole experiment. All children thus heard the exact same stimuli, the only difference between the two experiment versions being the button color assigned to *lut* (Version 1: green; Version 2: red) or *mip* (Version 1: red; Version 2: green).

Trial types. Children were exposed to three trial types. Two types were nonadjacent dependency utterances: *target items* ending in the target word (Version 1: *lut*; Version 2: *mip*), and therefore requiring a green button press; and *nontarget items* ending in the nontarget word (Version 1: *mip*; Version 2: *lut*), requiring a red button press. The third type were *filler items*, which did not contain a nonadjacent dependency as specified by the rule and required, similarly to the nontarget trials, a red button press because the last word was not the target (variable "*f*-element"; see below). Each trial (target, nontarget, or filler) consisted of three pseudowords with a 250-ms interstimulus interval between the three pseudowords. The average trial length was 2415 ms (min = 2067 ms; max = 2908 ms). Children had to press the button within 750 ms after the end of each utterance. If they did not do so, a null response was recorded and the next trial was delivered.

Eighty percent (216 trials) of the total 270 trials were target or nontarget trials. The structure of these trials was dependent on block type, as explained in the next section. The remaining 20% (54 trials) of all trials were fillers. The structure of these fillers was constant across the whole experiment and thus independent of block type. Fillers were built according to a *f* X *f* structure: 24 *f*-elements and 24 X-elements (Table 1) were combined under the constraint that the same *f*-element could not appear twice in the same utterance and that each X-element had the same probability to appear before or after a specific *f*-element. These fillers were added in anticipation of the disruption block, as explained in the next section.

Block types. There were three block types: training (3 blocks), disruption (1 block), and recovery (1 block). Each training block and recovery block consisted of 24 targets following one of the two nonadjacent dependencies (e.g., *tep* X *lut*), 24 nontargets following the other nonadjacent dependency (e.g., *sot* X *mip*), and 12 fillers. Each of the 24 unique target or nontarget trial combinations was

Table 1. *Overview of the 24 X-elements and 24 f-elements that were used to build the target, nontarget, 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

presented once per block, and repeated four times over the course of the whole experiment (three times in the training blocks and once in the recovery block). Unique filler item combinations were never repeated. This led to a total of 96 *tep* X *lut* trials, 96 *sot* X *mip* trials, and 48 *f* X *f* trials in the four sequenced blocks. The X-elements in the target or nontarget trials were selected from the same pool of 24 X-elements that was used for the filler items (Table 1).

The three training blocks were followed by one disruption block (30 trials). In this block, the (non)target did not comprise the nonadjacent dependencies presented in the training blocks. Instead their structure was *f* X *lut* and *f* X *mip*. *F*-elements and X-elements for these (non)targets were again selected from the elements presented in Table 1. Half of the X-elements were selected for the utterances with *lut* and the other half of the X-elements were selected for the utterances with *mip.* As a result, the disruption block had 12 *f* X *lut,* 12 *f* X *mip*, and 6 *f* X *f* trials. The disruption block was followed by the recovery block, which contained items structured similarly as the items in the three training blocks described above.

We predicted that if children are sensitive to the nonadjacent dependencies between each initial and final element in the target trials and nontarget trials, they should respond faster to target and nontarget items in the third training block and the recovery block compared to the disruption block (we will refer to this RT pattern as the *disruption peak*). Faster responses are expected in the third training block and recovery block as in these blocks, the initial word predicts the third word (and thus color of the button), whereas this is not the case in the disruption block in which all trials (target, nontarget, and filler) start with variable *f*-elements. By having filler items throughout the whole experiment, children are used to hearing utterances that start differently from *tep* or *sot*, ensuring that slower RTs in the disruption block are not simply a result of hearing utterances starting with novel pseudowords.

Offline measure of learning: Grammaticality judgments. After the recovery block, children received new instructions in which they were told that they would hear pairs of utterances and that they had to decide for each pair which of the two utterances was most familiar to the utterances in the previously heard language (e.g., *tep wadim lut* or *tep wadim mip*; two-alternative forced choice). In each utterance pair, one member followed the nonadjacent dependency rule (correct;

tep wadim lut in the example above) and the other member violated the non-adjacent dependency rule (incorrect; *tep wadim *mip* in the example above). Children were presented with 16 utterance pairs. In 8 of these utterance pairs, both members contained a novel X-element to test for generalization (*dufo*, *dieta*, *gopem*, *noeba*, *nukse*, *rolgo*, *sulep*, or *wiffel*). In addition, in each experiment version, half of the items assessed children's knowledge of their target NAD-rule (Version 1: *tep* X *lut*; Version 2: *sot* X *mip*) whereas the other half of the items assessed their knowledge of the nontarget NAD-rule (Version 1: *sot* X *mip*; Version 2: *tep* X *lut*). If needed, each single member of a pair could be repeated. Children had to respond verbally with "first" or "second." Their responses were recorded in E-prime by the experimenter.

Short debriefing: Awareness questionnaire. Once the children had completed all tasks, they were asked several questions regarding their awareness of the structures in the artificial language. Information concerning awareness of the non-adjacent dependencies is available for only half of the participants. The other half of the children received questions regarding their awareness of structure in a visual statistical learning task (see Procedure section and van Witteloostuijn, Lammertink, Boersma, Wijnen, & Rispens, 2017). Some of the questions included in this exit questionnaire aimed at gaining insight into participants' strategies during the exposure and grammaticality judgment phase (e.g., What did you focus on? Did you know when to press the green or red button or were you guessing?), while other questions directly asked whether participants had any explicit knowledge of the structure (e.g., complete the missing word in an utterance, did you notice a pattern and, if yes, explain what the pattern was).

Stimuli recording. All auditory stimuli were recorded in a sound-attenuated room by a female native speaker of standard Dutch. The stimuli were created following Gómez (2002), but slightly adapted to meet Dutch phonotactic constraints as in Kerkhoff, De Bree, and De Klerk (2013). The three-pseudoword-utterances featured a strong–weak metrical stress pattern, which is the dominant pattern in Dutch, and featured the following syllable structure: a monosyllabic word (*tep*, *sot*, or *f*-element) was followed by a bisyllabic word (X element), followed by a monosyllabic word (*lut*, *mip*, or *f*-element). The pseudowords were recorded in sample phrases and cross-spliced into the final utterances.

The auditory instruction given by *Appie the monkey* was recorded by a different female native speaker of standard Dutch. The speaker was instructed to use a lively and friendly voice as if she was voicing a monkey.

Procedure. All children performed three different tasks: the NAD-learning task (approx. 30 min), a self-paced visual statistical learning task (approx. 10 min), and a pilot version of a spelling task (approx. 5 min). In the current paper, we only report on the results of the NAD-learning task.

After every 30 utterances, children received feedback on the number of bananas they had picked (the monkey was awarded a banana whenever the child pressed the correct button). After the exposure phase, which lasted approximately

20 min, the children automatically received instructions on the grammaticality judgment task. This grammaticality judgment task was followed by an informal debriefing. For a visual representation of the word monitoring and grammaticality judgment tasks, see Figure 1.

Data preprocessing

Before analysing children's RT data and accuracy scores, the raw data set was preprocessed to remove unreliable measurements as described below.

Preprocessing RT data (online measurement). For the analysis of the RT data, all responses to filler items (20% of total trials) and all incorrect responses (17% of total trials[1]) were removed. RTs were measured from the onset of the third element and were considered an outlier whenever (a) children pressed a button before the onset of the third element or (b) when the RT for a particular trial type (target or nontarget) was 2 *SD* slower or faster than the mean RT for that particular trial type of the same child in the same block. A total of 256 (3.1%) outliers were removed. We used raw RT (instead of log-transformed RT data) as these are easier to interpret and both the quantile plot of the raw RT and the quantile plot of the log-transformed RT did not raise any concerns with respect to the normality of the residuals (see the Rmarkdown Main analyses script on our Open Science Framework project page: https://osf.io/bt8ug/). Finally, we selected children's RTs of the third training block, disruption block, and recovery block (4,464 observations) and used this data set to answer our research question.

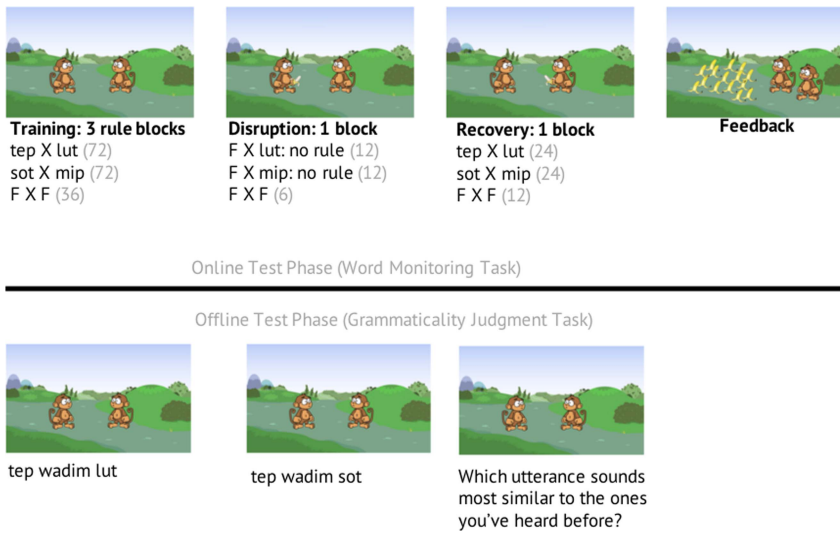

Figure 1. Visual representation of the online and offline test phases of the nonadjacent dependency (NAD)-learning task.

Preprocessing accuracy data (offline measurement). None of the responses in the grammaticality judgment task were removed. Responses were coded such that if the child picked the utterance with the trained nonadjacent dependency, the answer was judged correct (1), whereas the answer was judged incorrect if the child picked the utterance that violated the nonadjacent dependency rule (0).

Data analysis

RT data (online measure) were analysed using linear mixed-effects models (package *lme4*, Version 1.1-12; Bates, Maechler, Bolker, & Walker, 2015) in the statistical programming language *R* (R Core Team, 2017). For each relevant predictor, we computed its 95% confidence interval by the *profile* method; the corresponding *p* value was determined from the same profile iteratively (e.g., *p* was less than .05 if and only if the confidence interval did not contain zero). The dependent variable was RT as measured from the onset of the third element of the utterance. RT was fitted as a function of the ternary predictor Block (third training, disruption, or recovery), the binary predictors Targetness (nontarget or target) and ExpVersion (Version 1 or Version 2; see "Online Measures" in the Results section for more details), and the continuous predictor Age (in days). The predictors Block, Targetness, and ExpVersion were coded with sum-to-zero orthogonal contrasts (as detailed below) and the predictor age was centered and scaled. The RT model contained by-subject and by-item (X-element; $N = 24$) random intercepts, by-subject random slopes for the main effects of Targetness and Block as well as for the interaction between Targetness and Block, and a by-item random slope for ExpVersion.

Accuracy data of the grammaticality judgment task (offline measure) were analysed using a generalized linear mixed-effects model with accuracy (*correct* = 1; *incorrect* = 0) as the dependent variable. Accuracy was fitted as a function of the binary predictors Generalization (novel vs. familiar) and ExpVersion (Version 1 or Version 2) and the continuous predictor Age (in days). The binary predictors were coded with sum-to-zero orthogonal contrasts and the continuous predictor Age was centered and scaled. The accuracy model had by-subject and by-item (X-element; $N = 16$) random intercepts, by-subject random slopes for the main effects of Generalization, and a by-item random slope for ExpVersion.

Finally, we explored the relationship between children's online measure of learning (i.e., disruption) and their offline measure of learning. For each child, we computed an online disruption score by subtracting their average RT in the disruption block from their average combined RT in the third training block and recovery block combined. The proportion of correct answers on the grammaticality judgment task was taken as the offline measure of learning. The relationship between the online learning score and the offline learning score was explored with a Pearson's correlation coefficient.

In addition, we made our data, data preprocessing script (section 2.4), and analysis script available on our Open Science Framework project page: https://osf.io/bt8ug/. In the scripts on our Open Science Framework page, the reader can

also find the functions that we used to calculate *p* values and confidence intervals. Furthermore, on this page we provide the interested reader with some supplementary, exploratory descriptives and analyses that were requested by reviewers.

Predictions for the RT model (online measurement). As stated in our Materials and Procedure section, we predict that if children are sensitive to the nonadjacent dependencies they will show a disruption peak, meaning that RTs increase when the nonadjacent dependencies are temporarily removed (in the disruption block) compared to when the nonadjacent dependencies are present (in the third training block and the recovery block), for the target and nontarget items. Furthermore, we were interested in seeing whether this disruption peak is different for target items (requiring a positive response) versus nontarget items (requiring a negative response). Children's sensitivity to the nonadjacent dependency in target items might be different from their sensitivity to the nonadjacent dependency in the nontarget items for two reasons. First, the disruption peak in nontargets can be seen as a more indirect measure of sensitivity as nontarget items are less salient than the target items. Second, people are generally faster in giving a positive response (target items: green button) than a negative response (nontarget items; cf. López-Barroso et al., 2016). However, as exploring this difference was not part of our initial research question, it can be seen as a sanity check and therefore this analysis is exploratory (for more details see the Results section). We also check whether the disruption peak is different in experiment Version 1 (target: *lut*; nontarget: *mip*) from experiment Version 2 (target: *mip*; nontarget: *lut*), to check if counterbalancing yielded the desired results (viz. no evidence for a difference between experiment versions). Finally, we explored whether age modulates the size of the disruption peak.

Predictions for the accuracy model (offline measurement). For the accuracy measurement in the grammaticality judgment task, if children learn the nonadjacent dependency rules, their true mean accuracy scores on the two-alternative grammaticality judgment task (16 items) will exceed chance level. If children do not learn the nonadjacent dependencies, but rather recognize familiar items, their true mean scores for familiar items will be higher than those for novel items.

Prediction for the relationship between the online measurement and offline measurement. If we find a disruption peak, the relationship between children's online measure of learning and their offline measure of learning will be explored. In other words, it will be explored whether children that have a relatively large disruption peak also have a relatively high accuracy score on the offline grammaticality judgment task. As this comparison does not directly answer our research question, this analysis will be reported in the exploratory part of the results section.

Prediction for the awareness of nonadjacent dependencies. We predict that if children learn the nonadjacent dependencies explicitly (i.e., they can verbalize the nonadjacent dependency rule), they will be able to perform the sentence

completion task accurately in our short debriefing after the experiment and we hypothesize that they can verbalize the *tep* X *lut* and *sot* X *mip* dependency rules. For a summary of all confirmatory and exploratory hypotheses, see Table 2.

RESULTS

In this section, we distinguish between (a) descriptive results that are displayed for ease of exposition, (b) confirmatory results of our hypothesis testing, and (c) exploratory results that describe data checks and unexpected but interesting findings (cf. Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012). Note that in general one cannot draw any firm conclusions from exploratory results, so

Table 2. *Summary and operationalization of our confirmatory and exploratory research questions (RQs)*

Predictor	Operationalization	Type of RQ
Online RT measures		
Disruption peak	Do children's RTs increase when NAD-rules disappear?	Confirmatory
Targetness	Are children's RTs to target items (positive response) faster than their RTs to nontargets (negative responses)	Sanity check
Disruption peak × Targetness	Is the disruption peak different for target versus nontarget items?	Exploratory
Disruption peak × Age (days)	Does age modulate the size of the disruption peak?	Exploratory
Disruption peak × ExpVersion	Is there a difference in disruption peak between eperiment Version 1 and experiment Version 2?	Counter-balancing
Offline accuracy scores		
Intercept	Is children's accuracy score different from chance performance?	Confirmatory
Generalization	Is there a difference in accuracy between familiar and novel items?	Exploratory
Age (days)	Does age modulate children's accuracy score?	Exploratory
ExpVersion	Is there a difference in accuracy between experiment Version 1 and experiment Version 2?	Counter-balancing
Online disruption score and offline accuracy score		
Pearson's correlation	Are children's online disruption scores and their offline accuracy scores correlated?	Exploratory
Awareness questionnaire		
NA	Have children explicit knowledge of the nonadjacent dependency rules?	Exploratory

that only our confirmatory results can be used as evidence for the usability of RTs as an online measure of learning.

Online measure (RTs)

Online measure: Descriptives. Mean RTs to the target and nontarget items across the training blocks, disruption block, and recovery block are visualized in Figure 2. As we are interested in the learning trajectories across the third training block, disruption block, and recovery block, Table 3 lists the mean RTs with their residual standard deviation for these blocks only.

Online measure: Confirmatory results. To test our hypothesis of a disruption peak, we fitted a linear mixed-effects model restricted to the RTs of the third training block, the disruption block, and the recovery block (hereafter called the disruption model; Table 4). In order that our estimate of the effect of the first contrast ("Disruption Peak") of our ternary predictor Block represents the numerical height of the disruption peak in milliseconds, the coding of our sum-to-zero contrast for the ternary predictor has to contain a difference of 1: therefore the ternary contrast in the predictor Block ("Disruption Peak") estimated how much the true mean RT in the disruption block (which is coded as +2/3) exceeds the average of the true mean RT in the third training block (coded as –1/3) and the true mean RT in the recovery block (also coded as –1/3). This first contrast of the

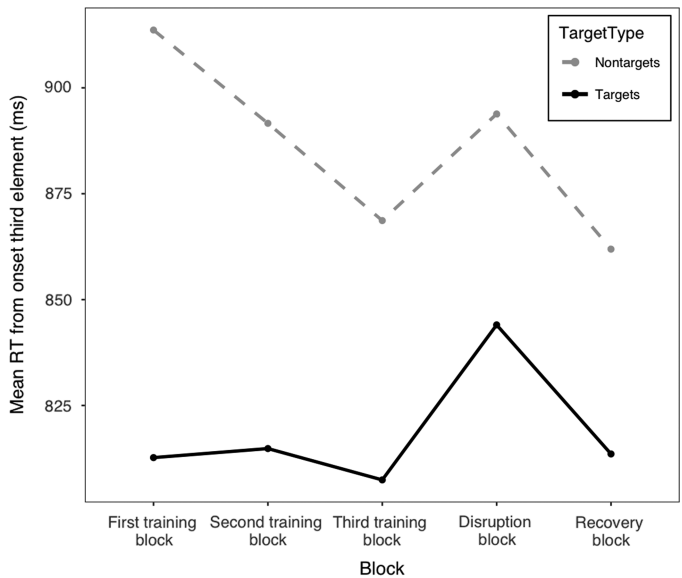

Figure 2. Mean response times to the target (black solid) and nontarget (gray dotted) items across the five blocks of exposure.

Table 3. *Response times in milliseconds to the target and nontarget items across the third training block, disruption block, and recovery block, separated by experiment version. Residual standard deviations (ms) as estimated by the linear mixed-effects model in parentheses*

Version 1 (target = *lut*)			
Trial type	Third training block	Disruption block	Recovery block
Target	749 (191)	777 (191)	761 (191)
Nontarget	842 (191)	869 (191)	836 (191)
Version 2 (target = *mip*)			
Target	875 (191)	921 (191)	875 (191)
Nontarget	900 (191)	921 (191)	891 (191)

predictor Block (Disruption Peak) intends to answer our specific research question (i.e., whether RTs are disrupted by removal of the nonadjacent dependency).[2] When we fitted the model, it showed a significantly positive effect of disruption peak. The disruption peak is 36 ms ($t = +3.8$; $p = .00038$; 95% CI [17, 56]). We thus conclude that children become 36 ms slower when we remove the nonadjacent dependency structure in target and nontargets items.

Online measure: Exploratory results. First, we checked that children, similarly to adults (López-Barroso, 2016), are faster in giving a positive than a negative response (Targetness). The model estimated that children's positive responses (average RT target items; +1/2) were 52 ms faster than their negative responses (average RT nontarget items; –1/2; $t = -4.6$; $p = .00003$; 95% CI [–75, –30]), so we can conclude that children are generally faster in giving a positive than a negative response.

Second, we explored whether the disruption peak differed between target items and nontarget items (interaction between Disruption Peak and Targetness). The disruption peak was 9 ms larger for target items than nontarget items, but not statistically significantly different from zero ($t = +0.55$; $p = .58$; 95% CI [–23, +41]). This means that we have no evidence that the height of the disruption peak differs between target items and nontargets items. To further explore this null result, we fitted two additional models in which we rereferenced the contrast coding. To obtain a t value for the disruption peak in target items, the contrasts were set as target 0 (previously –1/2) and nontarget +1 (previously +1/2). To obtain a t value for the disruption peak in nontarget items, the contrasts were set as target +1 and nontarget 0. For targets, the model estimated a disruption peak of 41 ms ($t = 2.9$; $p = .0042$; 95% CI [+13, +68]). For nontargets, the model estimated a disruption peak of 32 ms ($t = 2.7$; $p = .0068$; 95% CI [+9, +55]). Thus, both items types show a significant t value, suggesting that the disruption peak is present in both target items and nontarget items.

Table 4. *Outcomes of the disruption model (4,464 observations; N = 46); the last column (Relevance) explains how a certain comparison relates to our research questions, and if empty, the comparison is not of interest*

Random effects of subjects	*SD* (ms)
Intercept	80
Disruption peak	39
PrePostDisruption	25
Targetness	65
Disruption Peak × Targetness	53
PrePostDisruption × Targetness	27

Random effects of items (X-element)	*SD (ms)*
Intercept	92
ExpVersion	10

Fixed effect	*B* (*ms*)	CI_{low} (*ms*)	CI_{high} (*ms*)	*t*	*p*	*Relevance*
Intercept	+864	819	908	+39	7.4×10^{-9}	
Disruption peak	+36	+17	+56	+3.8	.00038	Confirmatory
PrePostDisruption	+1	–14	+15	+0.069	.94	
Targetness	–52	–75	–30	–4.6	3.0×10^{-5}	Sanity check
Age (days)	–13	–37	+11	–1.1	.29	
ExpVersion	+94	+45	+143	+3.8	.00034	
Disruption Peak × Targetness	+9	–23	+41	+0.55	.58	Exploratory
PrePostDisruption × Targetness	+13	–14	+40	+0.97	.33	
Disruption Peak × Age (days)	–5	–24	+14	–0.51	.61	Exploratory
PrePostDisruption × Age (days)	–5	–21	+10	–0.70	.48	
Targetness × Age (days)	+7	–16	+30	+0.60	.55	
Disruption Peak × ExpVersion	+6	–31	+43	+0.33	.75	Counterbalancing
PrePostDisruption × ExpVersion	–4	–34	+26	–0.25	.80	
Targetness × ExpVersion	+68	+22	+113	+3.0	.0043	
Age (days) × ExpVersion	–34	–83	+15	–1.4	.17	
Disruption Peak × Targetness × Age (days)	+21	–13	+55	+1.2	.22	
PrePostDisruption × Targetness × Age (days)	+8	–20	+35	+0.54	.59	
Disruption Peak × Targetness × ExpVersion	+37	–28	+102	+1.2	.25	

Table 4. *Continued*

Fixed effect	*B* (*ms*)	CI_{low} (*ms*)	CI_{high} (*ms*)	*t*	*p*	*Relevance*
PrePostDisruption × Targetness × ExpVersion	–6	–60	+47	–0.23	.82	
Disruption Peak × Age (days) × ExpVersion	+20	–18	+58	+1.1	.29	
PrePostDisruption × Age (days) × ExpVersion	+8	–23	+39	+0.52	.60	
Targetness × Age (days) × ExpVersion	+41	–5	+87	+1.8	.080	
Disruption Peak × Targetness × Age (days) × ExpVersion	–12	–81	+55	–0.36	.72	
PrePostDisruption × Targetness × Age (days) × ExpVersion	+10	–46	+65	+0.35	.73	

Third, we checked whether the disruption peak differs between the two versions of the experiment (interaction between Disruption Peak and ExpVersion). The model estimate of the interaction between DisruptionPeak and ExpVersion was not significantly different from zero (6 ms; $t = +0.33$; $p = .75$; 95% CI [–31, +43]). This null result for the counterbalancing interaction is good, as it means that we have no evidence that the size of the disruption peak differs between the two experiment versions and is thus dependent on the target dependency pair in focus. To further explore this null result, we again rereferenced the model contrasts to obtain a *t* value for the disruption peak in experiment Version 1 (ExpVersion 1: 0; ExpVersion 2: +1) and experiment Version 2 (ExpVersion 1: +1; ExpVersion2: 0). For experiment Version 1, the model estimated a disruption peak of 33 ms ($t = +2.5$; $p = .012$; 95% CI [+8, +59]). For experiment Version 2, the model estimated a disruption peak of 39 ms ($t = +2.9$; $p = .0054$; 95% CI [+12, +67]). In both experiment versions, the *t* value is significant, suggesting that the presence of a disruption peak is not dependent on the target dependency pair in focus.

Fourth and finally, we explored whether the size of the disruption peak is modulated by age (interaction between Age and Disruption peak). The model estimated that the disruption peak gets 5 ms smaller as children grow older, but this difference is not statistically significantly different from zero ($t = –0.51$; $p = .61$; 95% CI [–24, +14]). Thus, we have no evidence that the size of the disruption peak differs between younger and older children.

Offline measure (Accuracy grammaticality judgment)

Offline measure: Descriptives. Children's individual accuracy scores along with the overall mean accuracy score for the two-alternative grammaticality judgment task are visualized in Figure 3a. As a group, children selected the correct utterance with an accuracy of 51%, with individual accuracy scores ranging from 25% to 75%. As we also explore whether children scored better on familiar than novel items (Generalization), children's mean accuracy scores to these different item types are visualized in Figure 3b.

Offline measure: Confirmatory results. A generalized linear mixed-effects model was fit on the accuracy data of the offline two-alternative grammaticality judgment task to test whether children's accuracy scores on the task (16 items) exceeds chance level (736 observations; Figure 3a; Table 5). The predictor Generalization estimated by what ratio the children scored better on items with a familiar X-element (+1/2) than on items with a novel X-element (–1/2).

The model estimated that the children scored 1.6% above chance level (intercept: log odds +0.064, odds 1.07, probability 51.6%), but this was not statistically significant from chance ($z = +0.81$; $p = .42$; 95% CI [47.5, 55.7]). Therefore, we cannot conclude that learning of the nonadjacent dependencies can be evaluated via a two-alternative grammaticality judgment task. Furthermore, the model estimated that the children scored 0.90 times better (i.e., lower performance, as this odds ratio is less than 1) on novel (Generalization) than

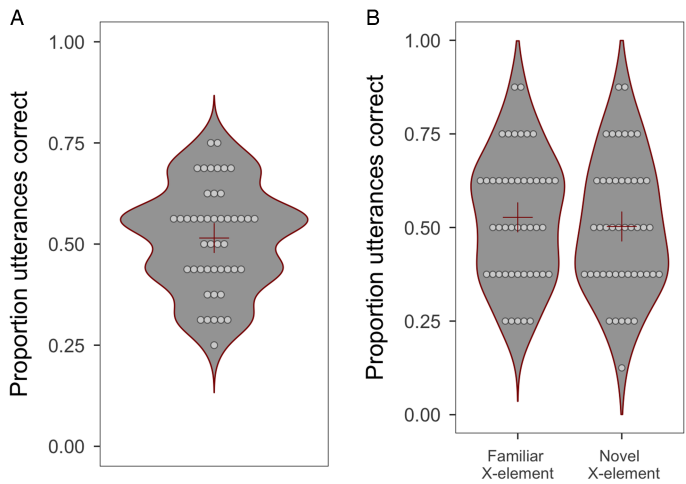

Figure 3. Violin plot that represents the distribution of (a) the overall mean accuracy scores on the two-alternative grammaticality judgment task and (b) the mean accuracy scores by generalization. The dots represent the individual scores, and the cross indicates the overall group mean.

Table 5. *Outcomes of the accuracy model (736 observations, N = 46); the last column (Relevance) explains how a certain comparison relates to our research questions*

Random effects of subjects							*SD (log-odds)*
Intercept							0.011
Generalization							0.063
Contrast	B_{model} (*log-odds*)	$B_{transformed}$ (*odds ratio*)	CI_{low} (*odds ratio*)	CI_{high} (*odds ratio*)	*z*	*p*	*Relevance*
Intercept	+0.064	1.07	0.90	1.26	+81	.42	Confirmatory
Generalization	–0.10	0.90	0.65	1.25	–66	.51	Confirmatory
Age (days)	–0.011	0.99	0.84	1.15	–0.14	.89	Exploratory
ExpVersion	–0.047	0.95	0.66	1.36	–0.27	.78	Counter-balancing
Generalization × Age (days)	+0.078	1.08	0.79	1.46	+0.51	.61	
Generalization × ExpVersion	–0.039	0.96	0.47	1.97	–0.11	.91	Counter-balancing
Age (days) × ExpVersion	–0.26	0.77	0.56	1.05	–1.68	.10	
Generalization × Age (days) × ExpVersion	0.15	1.16	0.63	2.16	+0.50	.62	

familiar items ($z = –0.66$; $p = .51$, 95% CI [0.65, 1.26]), but this ratio was not significantly different from 1 and therefore we cannot conclude that children treat novel items differently from familiar items.

Offline measure: Exploratory results. We checked whether children's accuracy scores were modulated by ExpVersion (counterbalancing; Version 1: –1/2; Version 2: +1/2) and Age. The model estimates of both predictors were not significantly different from 1 (Table 5), and therefore the results do not generalize to the population.

Relationship between online measure and offline measure of NAD-learning

Relationship between online measure and offline measure: Descriptives. For each child, we calculated an online disruption score and an offline learning score (Figure 4). Online disruption scores were computed by subtracting a child's average RT in the disruption block from his/her average combined RT in the third training block and recovery block (this is analogous to how the Disruption Peak contrast was calculated for the online measure). Hence, a positive outcome indicates that a child's RT in the disruption block was longer and thus slower than his/her combined average RT of the third training block and recovery block. Offline accuracy scores were obtained by calculating a child's proportion of correct answers on the offline grammaticality judgment task.

Relationship online measure and offline measure: Exploratory results. Pearson's correlation coefficient[3] between children's online measure of disruption and their

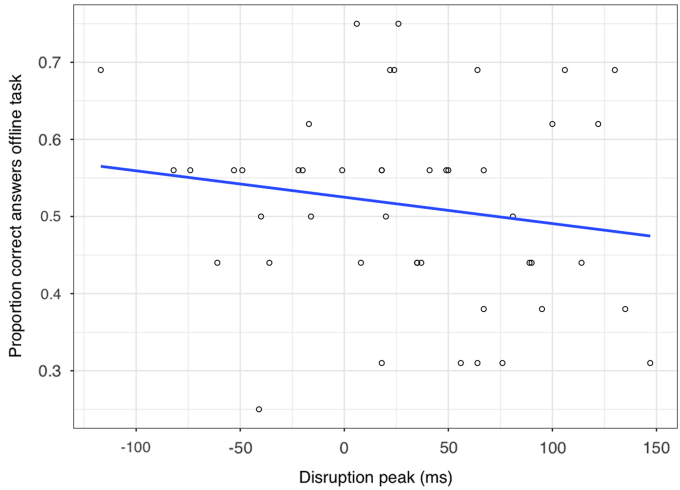

Figure 4. Scatter plot and regression line that represents the association between children's individual online disruption score (*x*-axis) and children's individual accuracy score on the grammaticality judgment task (*y*-axis).

offline measure of learning was not statistically significantly different from zero ($r = -.17$; $p = .27$; Figure 4). Therefore, we have no evidence that children's online disruption score correlates with their offline accuracy score.

Awareness questionnaire

None of the 24 children who were debriefed were able to verbalize either one or both of the nonadjacent dependency rules. In the sentence completion task, they were most likely to complete the utterance with the target word of the experiment version they were in. For example, a child who had to press the green button for *lut* (Version 1) replied *lut* to all the missing words in the sentence completion task, regardless of the missing words' positions and preceding or following words. We thus cannot conclude that children acquired any explicit (or at least verbalizable) knowledge of the nonadjacent dependency rules.

DISCUSSION

The current study was designed to investigate whether primary school-aged children are sensitive to nonadjacent dependencies in an artificial language and whether this sensitivity to nonadjacent dependencies could be measured (a) online by means of recording RTs and (b) offline by means of a two-alternative grammaticality judgment task, and (c) whether the online measure of sensitivity and offline measure of sensitivity were related to each other. Our results show that primary school-aged children are sensitive to nonadjacent dependencies in an

artificial language, at least in our online measure. As predicted, we found that when nonadjacent dependency rules were removed, the RTs increased relative to the RTs in the blocks that contained the dependency rules, indicating that children are sensitive to the nonadjacent dependencies.

The online measure can thus be seen as a promising advancement in measuring NAD-learning. On the basis of the offline measure alone, we would not have been able to conclude that children were sensitive to the nonadjacent dependencies (for similar findings in the SRT literature see Meulemans, Van der Linden, & Perruchet, 1998). It is important to note here, however, that we cannot directly compare our online measure and offline measure, and therefore, we would like to stress that we cannot conclude that online measures are better than offline measures (false p value comparison).

We like to speculate, however, that online and offline measures of nonadjacent dependency learning tap into different representations of acquired knowledge. This hypothesis has been proposed in previous studies on statistical learning in the auditory domain that also failed to find evidence of a relationship between the online and offline measures of learning (e.g., Franco et al., 2015; Isbilin et al., 2017; Misyak et al. 2010). In these studies, it is proposed that online measures are more sensitive to the transitional probabilities or co-occurrences present in the language whereas good performance on the grammaticality judgment tasks requires a comparison of two strings that can only be made from a more meta-linguistic or explicit decision (Franco et al., 2015). This meta-linguistic or explicit decision might be especially difficult for children as they acquire these skills relatively late. In addition, grammaticality judgment tasks similar to the one used in the current study have been argued to be psychometrically weak for measuring individual statistical learning performance (Siegelman et al., 2017). The latter raises the question as to how meaningful our exploration of the relationship between the online measure and offline measure of learning is. As we do believe that the online measure is an advancement, but not necessarily a substitute for the offline measure of nonadjacent dependency learning, we recommend that future studies try to improve the psychometric properties of the offline measures (for suggestions, see Siegelman et al., 2017) such that the online and offline measure of nonadjacent dependency learning are both informative as to whether children are sensitive to the nonadjacent dependency structure.

Furthermore, our exploratory finding that there is a disruption peak for both target and nontarget items suggests that the online measure of NAD-learning is not modulated by focus or saliency. One could argue that target items are more salient as they require a green button press. Therefore, a child may focus on hearing this target word while ignoring all other words. In addition, the target items (Version 1: *lut*; Version 2: *mip*) are explicitly mentioned during the instruction phase. Nontargets, by contrast, are not explicitly mentioned and therefore less salient than the targets items. Furthermore, as nontargets require a red button press, children might consider them as being less important. We have no evidence, however, that these differences in saliency do affect the size of the disruption peak. López-Barroso et al. (2016) report similar findings in their adult

version of the NAD-learning experiment. It is important to note that the word-monitoring task used in the current design *does* require a minimal level of attention to the stimuli, and therefore we cannot draw any conclusions on the specific incidental/implicit nature of NAD-learning with our task.

As discussed, the online measure of NAD-learning provides a promising advancement in measuring NAD-learning in typically developing primary school-aged children. Future studies could use the individual online disruption scores to further explore the relationship between children's sensitivity to nonadjacent dependencies and their sensitivity to (grammatical) structures in natural language. In adults, the online measure of sensitivity to nonadjacent dependencies is associated with adults' online processing (self-paced reading) of relative clauses such that better nonadjacent dependency learning is associated with faster processing of both subject relative clauses and object relative clauses (Misyak et al., 2010). We would be interested in seeing whether the same associations hold for typically developing children and whether we can take it one step further by investigating online nonadjacent dependency learning in children with language related impairments (developmental language disorder [DLD] and developmental dyslexia). The latter is of interest as statistical learning deficits have been proposed to explain parts of the language problems seen in people with a DLD (for meta-analytic reviews, see Lammertink, Boersma, Wijnen, & Rispens, 2017; Obeid, Brooks, Powers, Gillespie-Lynch, Lum, 2016). In these studies, we see that when people with DLD are compared to people without DLD, their offline grammaticality judgments are relatively poor. Similarly as for typically developing children, it could well be the case that people with a language disorder have difficulties explicitly judging grammaticality, resulting in lower offline judgment scores, not because they are worse learners, but simply because the task is too difficult or taps into a different type of acquired knowledge. Insight into the learning trajectories of both groups of learners could be beneficial and provide additional information on the statistical learning deficit in people with language impairments.

Finally, we believe that future (longitudinal) studies that aim to investigate the developmental trajectory of NAD-learning will benefit from the inclusion of our online measure of NAD-learning. Sensitivity to NADs can now be measured across all developmental stages (using different methods, as the current task is not feasible with infants; but see Cristia et al., 2016, for alternative measures of NAD-learning in infants). Capturing NAD-learning at different developmental stages is important as there is a vivid debate on the developmental trajectory of statistical learning (for reviews on this topic, see Arciuli, 2017; Krogh, Vlach, & Johnson, 2013; Zwart, Vissers, Kessels, & Maes, 2017).

Conclusion

In conclusion, this study was developed to obtain an online measure of statistical learning in children. RTs had already been shown to measure nonadjacent dependency learning in adults, and the applicability of this measure has now been extended to children.

ACKNOWLEDGEMENTS

This project is part of the project "Examining the contribution of procedural memory to grammar and literacy" awarded by the Dutch National Research Organisation (NWO) to Prof. Judith Rispens. We extend our gratitude to the four Dutch primary schools that participated. We also thank Darlene Keydeniers (test-assistant), Dirk Jan Vet (technical implementation of the experiment), and Josje Verhagen (advice on design experiment).

NOTES

1. Seventeen percent of target trials and 16% of nontarget trials, also the total percentage of errors, was approximately equally distributed across the five blocks.
2. The second contrast of Block ("PrePostDisruption") estimated how much the true mean RT in the recovery block (+1/2) exceeds the true mean RT in the third training block (–1/2). As this contrast does not directly answer our research question, we disregard the model outcome of this comparison.
3. Note that this correlation does not take into account the between-subject variable ExpVersion. In an alternative analysis, we added children's offline learning scores to the linear mixed effects disruption model (*OfflinePlus model*) and compared this OfflinePlus model to the disruption model (Table 4) by means of the analysis of variance function in R. When comparing both models, the OfflinePlus model did not significantly improve the disruption model ($\chi^2 = 1.74$; $p = .19$). Therefore, also when taking a slightly different approach that takes the between-subject variable ExpVersion and the random effects structure into account when comparing children's online disruption score and their offline accuracy score, we have no evidence that children's offline learning scores explain the variance in their online disruption scores.

REFERENCES

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

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bialystok, E. (1986). Factors in the growth of linguistic awareness. *Child Development*, *57*, 498–510.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy*, *21*, 648–667.

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

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

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. doi:10.1027/1618-3169a000295

Friederici, A. D., Mueller, J. L., & Oberecker, R. (2011). Precursors to natural grammar learning: Preliminary evidence from 4-month-old infants. *PLOS ONE*, *6*, e17920. doi:10.1371/journal.pone.0017920

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*, 183–206.

Iao, L. S., Ng, L. Y., Wong, A. M. Y., & Lee, O. T. (2017). Nonadjacent dependency learning in Cantonese-speaking children with and without a history of specific language impairment. *Journal of Speech, Language and Hearing Research*, *60*, 694–700. doi:10.1044/2016_JSLHR-L-15-0232

Isbilin, E., McCauley, S. M., Kidd, E., & Christiansen, M. (2017). Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. *Prooceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kerkhoff, A., De Bree, E., & De Klerk, M. (2013). Non-adjacent dependency learning in infants at familial risk of dyslexia. *Journal of Child Language*, *40*, 11–28. doi:10.1017/S0305000912000098

Krogh, L., Vlach, H. A., & Johnson, S. P. (2013). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology, 3*. doi:10.3389/fpsyg.2012.00598

Lammertink, I., Boermsa, 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. doi:10.1044/2017_JSLHR-L-16-0439

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. doi:10.1016/j.cognition.2016.03.016

McKercher, D., & Jaswal, V. (2012). Using judgment tasks to study language knowledge. In E. Hoff (Ed.), *Research methods in child language: A practical guide* (pp. 149–161). Oxford: Blackwell.

Meulemans, T., Van der Linden, M., & Perruchet, P. (1998). Implicit sequence learning in children. *Journal of Experimental Child Psychology*, *69*, 199–221. doi:10.1006/jecp.1998.2442

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

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162.

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

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

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1047–1052). Mahwah, NJ: Erlbaum.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Science*, *10*, 233–238.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Sandoval, M., & Gómez, R. (2013). The development of nonadjacent dependency learning in natural and artificial languages. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 511–522. doi:10.1002/wcs.1244

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2017). Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science. Advance online publication*. doi:10.1111/cogs.12556

Thomas, K. M., & Nelson, C. A. (2001). Serial reaction time learning in preschool- and school-age children. *Journal of Experimental Child Psychology*, *79*, 364–387. doi:10.1006/jecp.2000.2613

van Witteloostuijn, M., Lammertink, I., Boersma, P., Wijnen, F., & Rispens (2017). *Insights from novel measures of visual statistical learning in children.* Poster presented at the Congress of Interdisciplinary Advances in Statistical Learning. *Bilbao, Spain, June*, *28–30*, 2017.

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. doi:10.1177/1745691612463078

Zwart, F. S., Vissers, C. T. H., Kessels, R. P., & Maes, J. H. (2017). Procedural learning across the lifespan: A systematic review with implications for atypical development. *Journal of Neuropsychology.* Advance online publication. doi:10.111/jnp.12139