Detecting patterns

Relating statistical learning to language proficiency in children with and without developmental language disorder

It is still unclear why children diagnosed with developmental language disorder (DLD) experience so many difficulties acquiring their native language. The research described in this dissertation investigated whether differences in the ease with which children acquire language are related to children's sensitivity to statistical regularities (i.e. statistical learning) in the input. The following questions were addressed: (1) can we detect differences in statistical learning at the group and individual level (this concerns the measurement of statistical learning), (2) are individual differences in statistical learning associated with language proficiency and (3) can the language difficulties observed in children with DLD be explained by a statistical learning deficit that is observable across modalities, domains and dependency types?

With four empirical studies and two meta-analyses we aimed to answer these questions. Using online and offline measures of learning, we found evidence for statistical learning at the group level. Using these measures, we could not detect learning at the individual level (question 1). This means that we cannot draw a conclusion as to whether individual differences in statistical learning do (or do not) correlate with language proficiency (question 2). As for our third question: our results indicate that children with DLD have an auditory verbal statistical learning deficit. We cannot conclude that they have (or do not have) a statistical learning deficit outside this domain. The presence of a statistical learning deficit in children with DLD may thus depend on several factors, including the domain and modality in which learning is tested.

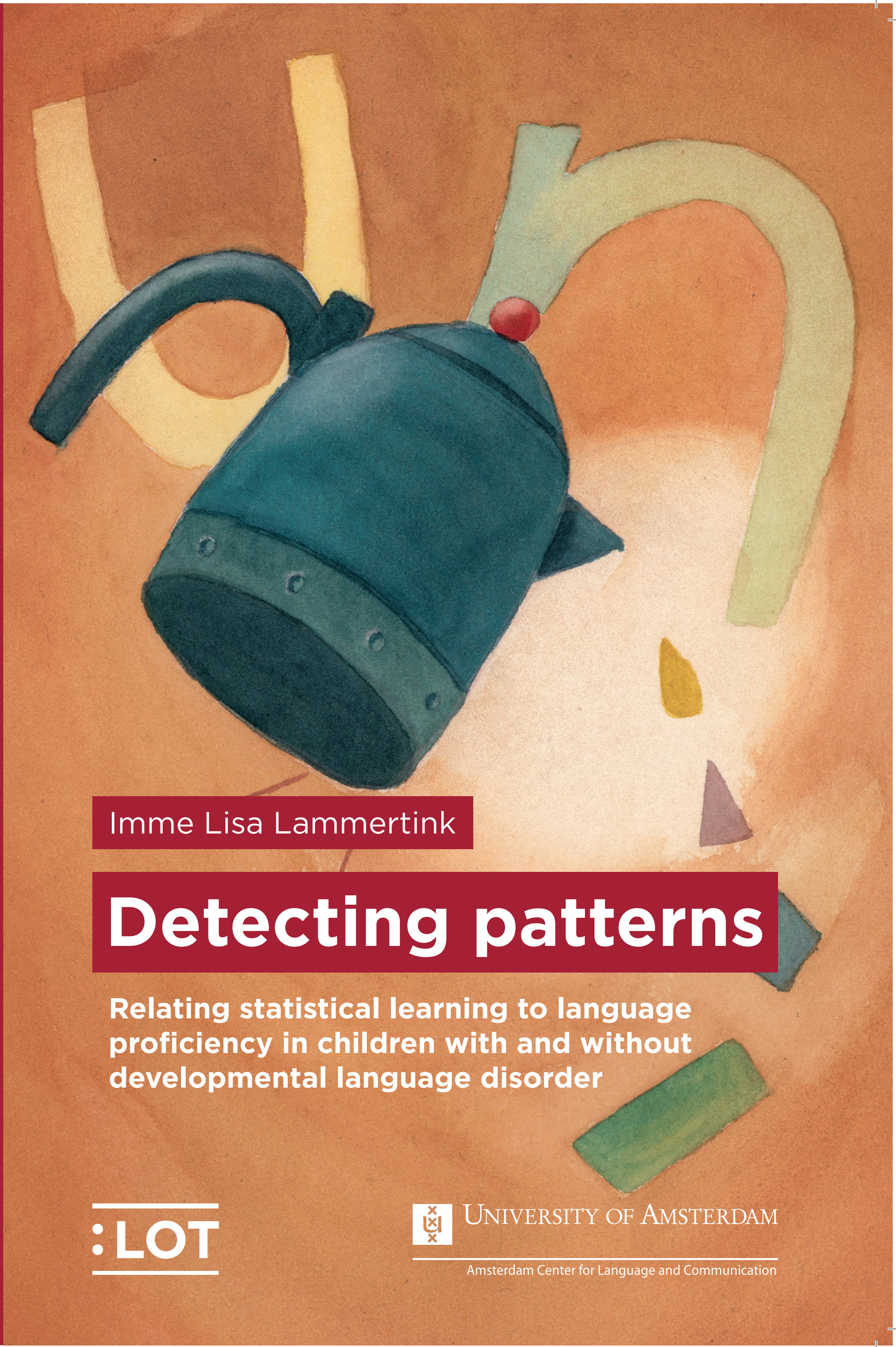
ISBN 978-94-6093-352-3

:LOT Netherlands Graduate School of Linguistics
Landelijke Onderzoekschool Taalwetenschap

:LOT
567

Imme Lisa Lammertink

Detecting patterns



Imme Lisa Lammertink

Detecting patterns

Relating statistical learning to language proficiency in children with and without developmental language disorder

:LOT

University of Amsterdam
Amsterdam Center for Language and Communication

Detecting patterns

Relating statistical learning to language proficiency
in children with and without developmental
language disorder

Published by
LOT
Kloveniersburgwal 48
1012 CX Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration © 2020: evelienjagtman.com. The percolator on the cover of this dissertation can be read as a metaphor: children with DLD had difficulties detecting linguistic patterns (symbolized by the letters ‘U’ and ‘N’ on the background), but when “poured’ in alternative form, they succeeded in their detection of non-linguistic visual patterns (symbolized by the percolator, which was built from these letters).

ISBN: 978-94-6093-352-3
NUR: 616

Detecting patterns: Relating statistical learning to language proficiency in children with and without developmental language disorder

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen
op donderdag 11 juni 2020 te 10:00 uur

door
Imme Lisa Lammertink
geboren te Nijmegen

Promotiecommissie:

Promotores:	prof. dr. P.P.G. Boersma	Universiteit van Amsterdam
	prof. dr. J.E. Rispens	Universiteit van Amsterdam
Copromotor:	prof. dr. F.N.K. Wijnen	Universiteit Utrecht
Overige leden:	prof. dr. R. de Diego Balaguer	University of Barcelona
	prof. dr. J. von Koss Torkildsen	University of Oslo
	prof. dr. A. Lukács	Budapest University of Technology and Economics
	prof. dr E.O. Aboh	Universiteit van Amsterdam
	prof. dr.J.C. Schaeffer	Universiteit van Amsterdam
	dr. J. Verhagen	Universiteit van Amsterdam

Faculteit der Geesteswetenschappen

This work is part of the research programme Vernieuwingsimpuls – Vidi (examining the contribution of procedural learning to grammar and literacy in children) with project number 27689005, which is (partly) financed by the Dutch Research Council (NWO).

In loving memory of my mother and Frits,
the “professors” of our family

Table of contents

Dankwoord (acknowledgements) **iii**
Author contributions **ix**

Prologue **1**

Chapter 1 General introduction **3**

Chapter 2 Statistical learning in specific language impairment: A meta-analysis **13**

Chapter 3 Auditory statistical learning in children: Novel insights from an online measure **37**

Chapter 4 Children with developmental language disorder have an auditory verbal statistical learning deficit: Evidence from an online measure **71**

Chapter 5 Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD **113**

Chapter 6 Statistical learning in the visuomotor domain and its relation to grammatical proficiency in children with and without developmental language disorder: A conceptual replication and meta-analysis **149**

Chapter 7 General discussion **181**

Epilogue **195**
References **197**
Appendices **221**
Summary in English **235**
Nederlandse samenvatting **241**
About the author **247**

Dankwoord (acknowledgements)

Toen mijn vader in 2016 naar Santiago de Compostella liep, was hij onder de indruk van de gastvrijheid waarmee hij overal ontvangen werd. Een soortgelijk gevoel heb ik tijdens het schrijven van dit proefschrift ervaren. “Onderweg” zijn er veel mensen geweest die mij, ieder op hun eigen manier, geholpen hebben.

Ten eerste wil ik het Progracy-team noemen. Toen “mijn boekje” af was, voelde dat heel dubbel. Met dit team had ik eigenlijk nog wel jaren door willen gaan. **Judith**, dank voor alle vrijheid die je me gaf. Van tevoren wist je niet zeker of ik nog wel terug zou komen uit Melbourne of Den Haag, maar beide keren besefte ik met hoeveel plezier ik aan ons Progracy-project werkte. En beide keren keek ik ernaar uit weer fulltime met Progracy aan de slag te gaan. Dank dus voor het vertrouwen dat je me gaf, maar vooral ook voor het bedenken van dit project en voor je interesse in mij als persoon. **Paul,** jou wil ik bedanken voor je kritische blik en oog voor detail. Je leerde me om nooit zomaar iets voor waar aan te nemen, wat ik als heel waardevol ervaar. **Frank**, waar ik soms het gevoel had dat we met het isoleren van een statistisch leermechanisme de complexe werkelijkheid nooit zouden begrijpen, overtuigde jij mij dat dit soort fundamenteel onderzoek nodig is om dat uiteindelijk wel te kunnen. **Merel**, mijn *partner in crime*! Het was ontzettend fijn samenwerken en ik heb veel van je geleerd. Maar misschien wel het belangrijkste, dankzij jou waren de afgelopen vierenhalf jaar een stuk gezelliger!

Ruth de Diego Balaguer, Janne von Koss Torkildsen, Ágnes Lukács, Enoch Aboh, Jeannette Schaeffer and **Josje Verhagen,** I am honoured that you agreed to be part of my committee and I am looking forward to discuss the outcomes of my studies with you during my defence.

Uiteraard wil ik ook **alle kinderen** die meegedaan hebben aan mijn onderzoek bedanken. Zonder hen was dit proefschrift er niet geweest – de enige conclusie die ik na vierenhalf jaar onderzoek met 100% zekerheid kan trekken. Ook wil ik

alle ouders en **scholen** bedanken voor hun gastvrijheid. Van de gesprekken die ik met ouders en leerkrachten heb gevoerd, heb ik ontzettend veel over TOS geleerd.

Het gevoel dat mijn vader ervaarde tijdens zijn wandeltocht heeft waarschijnlijk de grootste gelijkenis met de hulp die ik kreeg van alle **logopedisten**, **klinisch linguïsten**, **ambulant begeleiders**, **leerkrachten**, **teamleiders** en **schooldirecteuren** werkzaam bij **Auris**, **Kentalis, Pento** en **Viertaal** tijdens het werven van deelnemers. In het bijzonder wil ik **Erik Hersbach**, **Lusanne van der Velden** en hun collega's van **Viertaal Schagen** bedanken. In Schagen mocht ik twee jaar op rij testen en beide keren was ik echt even onderdeel van het team, dank hiervoor! Ook wil ik **Margo Zwitserlood** van **Pento** persoonlijk noemen. Jouw hulp leverde me aan het eind van de wervingsperiode onverwacht heel wat extra kinderen op. Dank ook aan **Arend Verschoor** van **FOSS/ Stichting Hoormij** voor het meermaals delen van mijn wervingsbrochure.

Ellen, Darlene, Iris, Maartje en **Sascha**, jullie hulp was onmisbaar bij het verzamelen en coderen van de controledata (van zowel volwassenen als kinderen!). **Iris,** jou wil ik in het bijzonder bedanken voor het afreizen naar de meest onbereikbare plekken in Nederland om data voor mij te verzamelen. En wat was ik blij toen je als PhD bij ons kwam werken! **Iris, Sascha, Sybren, Thomas** en **Veronika** dank voor het schrijven van interessante scripties binnen ons project. **Sybren,** alvast bedankt voor het lezen van mijn volledige proefschrift. Het is fijn te weten dat tenminste één iemand dat zal doen ;).

In de afgelopen jaren ben ik vaak van Nijmegen naar Amsterdam gereisd. Deze reis maakte ik graag, omdat ik wist dat er in Amsterdam altijd wel collega-promovendi aanwezig zouden zijn. Lieve **Cindy, Iris, Marieke, Marloes, Merel** en **Ulrika,** ik kan niet vaak genoeg zeggen hoeveel ik de lunches en wandelingetjes naar "De Italiaan" ga missen. **Hernán** en **Natalia,** dank voor de leuke lees, - en schrijfgroep bijeenkomsten. En met veel plezier denk ik terug aan de tijd dat alle bureaus in PCH 6.50 nog gevuld waren. **Iris B**, **Iris D**, **Merel, Tessa** en **Tiffany**, dank voor deze fijne tijd. Ik hoop dat we elkaar blijven zien.

Ik wil ook alle onderzoekers van **Grammar & Cognition** en het **ACLC** bedanken. De meetings en lezingen op vrijdagmiddag waren een extra stimulans om in de trein te stappen. Ook **Dirk-Jan Vet** heel erg bedankt: niet-werkende E-prime-scripts zijn heel frustrerend, maar zo heb ik onze afspraken nooit ervaren. Dank voor je engelengeduld en voor je oprechte interesse in mijn onderzoek.

Ook in Nijmegen wil ik een aantal (oud-)collega's bedanken. **Paula**, **Caroline** and all the others members of the **FLaDD** group, thanks for having me as a guest in your Monday meetings. **Marisa**, I am happy that I accepted your challenge to create our own language. Nepperlands always reminds me of how much fun science can be: we bejen niet op, het lort heel mooi! **Antje**, thanks for always hosting me whenever I am around. I really enjoy visiting you in so many different places.

I would also like to thank **Jarrad Lum** for having me at Deakin University, Melbourne. Running a brain stimulation study is, beyond doubt, one of the most exciting things I ever did.

In het laatste jaar van mijn promotie heb ik zes maanden als junior beleidsmedewerker bij NWO gewerkt. Hier heb ik geleerd dat het ook buiten de universiteit hard werken is, maar tegelijkertijd heel gezellig kan zijn. Dank aan alle collega's van **SGW** voor deze ervaring. **Joyce** en **Jeroen**, ook jullie wil ik bedanken. Het was ontzettend fijn om in Den Haag zo'n mooi thuis te hebben!

Op verschillende manieren heb ik de afgelopen jaren de vertaalslag van taalwetenschappelijk onderzoek naar de maatschappij en praktijk mogen maken. Dit vond ik een van de leukste dingen om te doen. In dit kader wil ik **Darlene, Iris, Jael, Klaske, Linda, Nina, Saskia** en **Tessa** – mijn (oud-)WAP-redactiegenoten – danken voor de inspirerende Bulletins die we samen maken. Dank ook aan **Akke** met wie ik regelmatig brainstormde over de WAP-website, altijd onder het genot van een warme kop soep. **Karin** en **Willemien**, het is een hele eer om via **Wetenschap.nu** mijn fascinatie voor taalverwerving te mogen delen. **Suzanne** en **René**, dank voor **Sciencebattle,** een prachtig concept om meer aandacht voor wetenschappelijk onderzoek (en in mijn geval TOS) te genereren.

Daan en **Evelien**, dank dat jullie met PIEWOLC op mij hebben gewacht en dat we samen zo'n mooie uitdaging aan mogen gaan. Dank ook aan mijn nieuwe collega's bij **Kentalis** en **NSDSK** voor de warme ontvangst. De leegte die ik ervaarde na het indienen van mijn manuscript was hiermee snel gevuld.

Naast collega's wil ik ook graag een aantal vrienden bedanken die ieder op hun eigen manier (waarschijnlijk ongemerkt) een steentje hebben bijgedragen. **Bernard, Ellen, Elske** en **Linda,** soms is het fijn om gewoon even je verstand op nul te kunnen zetten. Jullie weten als geen ander hoe dat (al fietsend of lopend) moet! En **Elske**, de vele muhammara's en sinaasappel-gembertheetjes waren altijd een welkome afleiding op mijn thuiswerkdagen! **Anne**, **Erin** en **Natasja**, dank voor de gezellige etentjes, bruiloften en babyshowers. Jullie helpen me altijd herinneren dat er meer is in het leven dan onderzoek (en sport)! **Renske** en **Rutger**, het was fijn dat ik altijd bij jullie in Haarlem terechtkon en dat er dan ook nog eens een heerlijke maaltijd voor me klaarstond. Wat een luxe! **Ria**, ontzettend bedankt voor de bijzondere band die we de afgelopen jaren hebben opgebouwd. **Yvonne**, bedankt voor het delen van mooie herinneringen. Heel waardevol in een fase waarin ik me vooral op de toekomst richt.

Marloes en **Tessa**, ik ben vereerd dat jullie mijn paranimfen willen zijn. **Lieve Marloes**, stiekem denk ik dat we best een beetje trots mogen zijn op "onze" Research Fair. Mede dankzij de organisatie van dit evenement veranderde onze collegiale band in een vriendschap. Deze vriendschap werd voor mijn gevoel nog sterker toen ook jij "voor de spits" ging reizen en we de dag af en toe samen begonnen met een goed ontbijt. Die ontbijtjes mis ik nu al. **Lieve Tes,** van penvriendin tot paranimf, dat is onze vriendschap in een notendop. We ontmoetten elkaar in Frankrijk toen we de leeftijd hadden van de kinderen die ik getest heb. Ik ben ontzettend blij dat we na deze vakantie bleven schrijven. Het voelt heel bijzonder om zo'n sterke vriendin te hebben, op wie ik altijd kan rekenen en met wie ik alles kan delen.

In de afgelopen vierenhalf jaar heb ik ook geleerd hoe belangrijk het hebben van een lieve (schoon)familie is. **Leendert**, **Liesbeth**, **Otto** en **Giny**, jullie weten als geen ander hoe het is om een boek te schrijven en de publicatie daarvan feestelijk te vieren. Dank voor dit goede voorbeeld én dank voor het hebben van contacten wereldwijd die mij tijdens werkbezoeken op sleeptouw kunnen nemen. **Hanke**, **Jan** en **Miriam**, dank voor jullie interesse in mijn werk en voor het spelen van proefpersoon ☺.

Lieve tante Anna, het schrijven van een proefschrift vergt heel wat doorzettingsvermogen en van tijd tot tijd ook het vermogen om met tegenslagen om te gaan. In beide aspecten ben jij een voorbeeld. Ik wil jou vragen om dit boek ook namens mama, Marika en Frits in ontvangst te nemen.

Lieve Papa en Jacintha, bij jullie ben ik altijd welkom. Dank voor het vertrouwen dat jullie in mij hebben. Dit vertrouwen heb ik nodig om in mijzelf te blijven geloven. Papa, dank ook voor het zijn van een grote inspiratiebron én voor al die uren klussen in ons nieuwe huis zodat ik de laatste hoofdstukken van dit proefschrift kon schrijven ;).

Lieve Len, wij hebben geen woorden nodig om elkaar te begrijpen, iets wat een taalwetenschapper niet snel zal zeggen. Je hebt de gave om altijd op het juiste moment te informeren hoe het met me gaat, geeft de beste adviezen en leerde mij dat niets onmogelijk is!

Last but not least, **lieve Teije**. Veel van mijn doelen had ik niet zonder jou kunnen bereiken. Letterlijk en figuurlijk help je mij grenzen te verleggen. Zonder jouw vertrouwen, geduld, liefde en steun was dit boek er nooit gekomen. Ik hoop dat we samen nog veel mogen ontdekken.

Author contributions

All the research described in this dissertation is based on the project proposal Examining the Contribution of Procedural Learning to Grammar and Literacy Acquisition in Children written by Judith Rispens (JR). JR received a VIDI grant, awarded by the Dutch Research Council (NWO), for this proposal in 2014.

Chapters 1 and 7

Imme Lammertink (IL) wrote these chapters and made revisions based on feedback provided by Paul Boersma (PB), Frank Wijnen (FW) and JR.

Chapters 2, 3, 4, 5 and 6: statistical analyses

IL and PB did the statistical analyses. IL wrote the scripts for analyses and made them publicly available via the Open Science Framework project pages. For chapters 3, 4, 5 and 6 PB wrote the *get.p.value* function and for chapter 5 he wrote the *get.p.value.3AFC* function.

Chapters 2, 3, 4, 5 and 6: writing

IL wrote a first version of each of these five chapters. These first versions were discussed with PB, FW, and JR during supervision meetings and via email. On the basis of these discussions, IL revised the chapters into versions that were submitted for publication. During the review process that followed (chapters 2, 3, 4 and 5), IL revised the chapters on the basis of feedback from anonymous reviewers. These revisions were again discussed with PB, FW and JR and resulted in published versions of chapters 2, 3, 4 and 5. For the chapter versions printed in this dissertation, IL made minor changes to text and formatting.

Chapters 4, 5 and 6: recruitment, testing and scoring

IL recruited the children with developmental language disorder (DLD). IL, Merel van Witteloostuijn (MvW) and research assistants recruited the typically developing children. IL and MvW trained and supervised the research assistants in testing the children and in scoring the data. Research assistants tested the typically developing children and scored their data. IL and research assistants tested the children with DLD and scored their data.

Chapter 2

This chapter is a slightly modified version of the paper that was published as: 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.

Data and scripts for analyses: https://osf.io/4exbz/

IL and JR defined the search term (see appendix A2.1) and inclusion criteria for the meta-analysis. IL and a research assistant conducted the literature search. IL created the community-augmented database that comes with this chapter (the use of such community-augmented database is based on Tsuji, Bergmann and Cristia, 2014).

Chapter 3

This chapter is a slightly modified version of the paper that was published as: Lammertink, I., van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019). Auditory statistical learning in children: Novel insights from an online measure. *Applied Psycholinguistics*, *40*(2). 279–302.

Data and scripts for analyses: https://osf.io/bt8ug/

IL, MvW and JR designed the auditory verbal nonadjacent dependency learning task. The design of the task is based on López-Barroso, Cucurell, Rodrīguez-Fornells, & de Diego-Balaguer (2016), but modified into a child-friendly version. Dirk Jan Vet (DJV) helped with the technical implementation of the task in E-prime. The task instructions were written and recorded by IL. The nonsense words were constructed by IL (based on Gómez, 2002 and Kerkhoff, de Bree, de Klerk and Wijnen, 2013) and recorded by JR.

Chapter 4

This chapter is a slightly modified version of the paper that was published as: Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Children with developmental language disorder have an auditory verbal statistical learning deficit: evidence from an online measure. *Language Learning*, *70*(1), 137–178.

Data, materials and scripts for analyses: https://osf.io/8a3yv/

The task used in this chapter is a slightly modified version of the task that was described in Chapter 3. IL, MvW and JR decided on these adaptations and DJV implemented them in the E-prime script

Chapter 5

This chapter is a slightly modified version of the paper that was accepted for publication as: Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing: An Interdisciplinary Journal*. Advance online publication.

Data and scripts for analyses: https://osf.io/8gpjt/

MvW, IL and JR designed the visual statistical learning task. The design of the task is modeled after tasks described by Arciuli and Simpson (2012), Siegelman, Bogaerts, and Frost (2017), Siegelman, Bogaerts, Kronenfeld, and Frost (2018) and is almost identical to the visual statistical learning task as described in van Witteloostuijn, Lammertink, Boersma, Wijnen, & Rispens (2019). DJV helped with the technical implementation of the task in E-prime. MvW created the alien cartoons and corresponding triplets. MvW also wrote the task instructions.

Chapter 6

This chapter is a slightly modified version of the paper that is under review as: Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (under review). Statistical learning in the visuomotor domain and its relation to grammatical proficiency in children with and without DLD: A conceptual replication and meta-analysis.

Data and scripts for analyses: https://osf.io/e9w43/

Chapter 6 consists of two parts: an experiment and a meta-analysis. The serial reaction time task that was used in the experimental part of this chapter was designed by Evan Kidd and Jarrad Lum (Lum & Kidd, 2012). DJV, MvW and IL implemented a Dutch version of this task in E-prime (using the exact same script, but with both the audio and written instructions translated into Dutch). The audio instructions were recorded by a student assistant.

For the meta-analysis, IL defined the search term (see appendix A6.1) and inclusion criteria. IL and two research assistants conducted the literature searches. IL created the community-augmented database (Tsuji et al. 2014) and entered the relevant data. One research assistant checked the data that IL entered into the database.

Prologue

When asked to explain the research described in this book, I usually quote a passage from Dave Eggers' roman *The Circle.* In this passage, the main character of the book (Mae) meets a little boy, around three years old, called Michael. Crucially, Michael wears a silver watch that recognizes, categorizes and counts the words spoken to him. It is explained to Mae that this counting of words is important as "studies show that kids need to hear at least 30.000 words a day" (Eggers, 2013, p.338). Though I never verified the truth of these 30.000 words, the example nicely illustrates that it is common knowledge that children need language input to learn their language from. Different views exist on what kind of learning mechanism children use to learn language from the input, however. One perspective is that already during the earliest stages of language development, children unconsciously detect and extract regularities (statistical patterns) from their language input that reflect the possible sound combinations, words and grammar of their native language. The research described in this book aims to investigate if children indeed detect such regularities and whether the ability to do so is associated with language proficiency.

Chapter 1

1.1 General introduction

To become a proficient language user, young children need to learn how their language is structured at the sound, word and sentence level. The language input that children receive may facilitate this process as the input is rich in terms of statistical regularities that reflect the linguistic structure (Reali & Christiansen, 2005). For example, in the English present tense, singular subjects frequently co-occur with [s] marking on the verb, whereas such marking is absent in the case of plural subjects (subject verb-agreement: *the child* walk*s* versus *the children* walk). As children grow older, they receive more linguistic input and will thus encounter more instances of these singular subject plus verb-plus-[s]. Importantly, the singular subject + [s] marking is constant whereas the verb varies (e.g., *he* walk*s*, *he* talk*s*, *he* eat*s*). This variability in verbs makes the marking more salient and easier to detect. One theory on first language acquisition that has been proposed is that children have a general (non-language-specific) cognitive capacity that is sensitive to such structural regularities in a variety of input. This cognitive ability is referred to as “statistical learning” (Saffran and Kirkham, 2018).

Children are likely to be different in their sensitivity to statistical regularities in the environment. For some children it may be more difficult to detect statistical regularities, and these children may need more or different input (Plante & Gómez, 2018) than other children who have fewer difficulties detecting the regularities. In this dissertation it is investigated (1) whether we can detect such differences in statistical learning ability at the group and individual level (this concerns the measurement of statistical learning), (2) whether these individual differences in statistical learning ability correlate with language proficiency and (3) whether the problems observed in children with a language disorder (developmental language disorder, explained later on) can be explained by a statistical learning deficit (as for example proposed by: Evans, Saffran, & Robbe-Torres, 2009; Hsu & Bishop, 2014a; Obeid, Brooks, Gillespie-Lynch, & Lum, 2016; Wijnen, 2013).

A central debate within the study on statistical learning and its relation to language proficiency concerns the specificity of the mechanism. In the examples given so far, the focus was on linguistic input. However, structure is not unique to human language. For example, music and bird songs also contain structural regularities (Rohrmeier, Zuidema, Wiggins, & Scharff, 2015). Therefore, it may well be the case that children use a statistical learning mechanism that is sensitive to regularities across domains (verbal, nonverbal) and modalities (auditory, visual, visuomotor) rather than a language-specific statistical learning mechanism in language learning. As the studies reported in this dissertation investigate statistical learning and the presence of a statistical learning deficit across domains and modalities (see Outline and research questions), this dissertation contributes to the scientific debate on the specificity of statistical learning and the statistical learning deficit (for a recent review discussing the specificity of statistical learning, see Frost, Armstrong, & Christiansen, 2019).

1.1.1 Statistical learning as a mechanism involved in language acquisition

Central to the studies described in this dissertation is the hypothesis that children learn certain aspects of language with a general cognitive (non-language-specific) learning mechanism that is capable of detecting statistical patterns in a broad variety of stimuli. In other words, children may learn language with a mechanism that also supports learning of nonverbal structure. Not all theories of language acquisition, however, agree that such a general non-language-specific mechanism plays a critical role in language learning. It has also been argued that children learn language via devices that are specialized in doing so (e.g., Chomsky, 1965; Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Lenneberg, 1967; Pinker, 1994). Chomsky (1965) for example, argues that children are born with a set of innately specified possible linguistic structures and that children may use a specialized "language acquisition device" to select the appropriate set of structures, i.e., those that represent the grammar of their native language, from the linguistic input. Central in Chomsky's reasoning that children must have innately specified linguistic structures is the "poverty of the stimulus" argument: children cannot induce (or generalize) language structure solely from their input, because the input they receive is restricted. That is, certain linguistic structures that do exist in the child's language may never (or hardly ever) occur in the child's input, making it impossible to learn these structures via induction. Over the past few years, several experimental, computational and corpus studies provided evidence against the poverty of the stimulus argument (Pullum & Scholz, 2002). These studies showed that children can use the rich statistical structure of their linguistic input to make inferences about the acceptability of structures that they have not encountered in their input yet (e.g., Reali & Christiansen, 2009 argue that this is the case for auxiliary fronting of polar interrogatives). At this point it may be good to highlight that both approaches to language acquisition acknowledge that input plays a significant role in learning language (Lidz & Gagliardi, 2015). The approaches differ as to (a) the linguistic nature of the learning mechanism involved in language learning and (b) the amount of innately specified (abstract) linguistic knowledge.

That infants are sensitive to statistical regularities in the linguistic input was first shown by Saffran, Newport and Aslin (1996). In this study, 8-month-old infants were able to use statistical information (i.e. transitional probabilities) to discover word boundaries in a continuous stream of auditorily presented nonsense syllables. This type of statistical learning is often referred to as sequential statistical learning, that is sensitivity to the ordering or co-occurrence of elements (segments, syllables, morphemes, words) over time. People are also sensitive to other types of statistical information, such as distributional cues and cross-situational dependencies. Evidence for infants' sensitivity to distributional statistics comes (amongst others) from Maye, Werker and Gerken (2002). In their study Maye et al. (2002) exposed infants to novel speech sounds. The speech sounds were arranged according to either a bimodal distribution or a unimodal distribution. It was hypothesized that if the infants were sensitive to the distributional statistics, the infants from the bimodal condition should form two-category representations (they should distinguish [d] and [t]) whereas the infants from the unimodal condition should form one-category representations. Maye et al. (2002) concluded that this hypothesis was confirmed and thus that infants are sensitive to distributional statistics. The third type of statistical learning, cross-situational statistical learning is often investigated in the context of lexical learning. For example, Yu and Smith (2007) and Smith and Yu (2008) showed that both adults (Yu a& Smith, 2007) and infants (Smith & Yu, 2008) compute distributional statistics across the co-occurrence of words and referents over multiple trials. That is, in cross-situational statistical learning studies participants are presented with individual trials that consist of a label and multiple pictures representing possible referents. Based on one trial it is not possible to connect the label with a referent. Across trials the specific word-referent mappings are consistent; a word of a specific word-referent pair is only used if the accompanied referent is also present. Therefore, participants can learn the specific word-referent mappings via a mechanism that is sensitive to the co-occurrences of specific word-referent pairs across trials. All studies in the present dissertation focus on sequential statistical learning, and therefore when the term statistical learning is used, it refers to sequential statistical learning (unless explicitly specified differently).

As mentioned before, statistical regularities also occur outside the linguistic domain and studies have shown that people are sensitive to regularities in these other domains as well. For example, Saffran, Johnson, Newport and Aslin (1999) showed that, using 12 tones from a musical octave, 8-month-old infants discriminated statistically coherent tone-triplets from slightly less coherent tone-triplets. Beyond the auditory modality, sensitivity to nonverbal sequences in the visuomotor domain is commonly observed with the serial reaction time task (e.g., Meulemans, van der Linden, & Perruchet, 1998; Nissen & Bullemer, 1987; Thomas & Nelson, 2001). In the visual domain, there is evidence that infants, children and adults use transitional probabilities of occurring visual shapes (abstract shapes or cartoon drawings) to form boundaries between pairs or triplets of visual elements (e.g., Arciuli & Simpson, 2011; Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002; Turk-Browne, Jungé, & Scholl, 2005).

It has also been shown that sensitivity to these regularities, both in the verbal and nonverbal domain, is associated with language proficiency. This evidence comes from two different sources. Firstly, children with atypical language development (dyslexia; developmental language disorder) may have a statistical learning deficit (for a review see Arciuli & Conway, 2018). Secondly, at the individual level there is evidence that statistical learning ability correlates with language proficiency. That is, better statistical learners have larger vocabularies (Spencer, Kaschak, Jones, & Lonigan, 2015; Shafto, Conway, Field, & Houston, 2013) and score better on tasks that tap into grammatical proficiency (Hamrick, Lum, & Ullman, 2018), syntactic processing (Kidd, 2012; Kidd & Arciuli, 2016; Wilson et al., 2018) and reading (Arciuli & Simpson, 2012; Hedenius et al., 2013; Steacy et al., 2019; Vakil, Lowe, & Goldfus, 2015; van der Kleij, Groen, Segers, & Verhoeven, 2018; von Koss Torkildsen, Arciuli, & Wie, 2019). It is important to note, however, that different research groups have raised their concerns about the existence of a publication bias in the literature on statistical learning deficits (Schmalz, Altoè, & Mulatti, 2017; van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017) and on the use of psychometrically weak measures of individual measures of statistical learning (Arnon, 2019; Siegelman, Bogaerts, & Frost, 2017; West, Vadillo, Shanks, & Hulme, 2017). These issues may inflate the conclusions drawn so far and are also repeatedly discussed in the individual chapters of this dissertation.

1.1.2 Developmental language disorder and the statistical learning deficit hypothesis

Central to this dissertation are children with developmental language disorder (DLD). These children experience difficulties with language across multiple areas such as the lexicon, morphology, (morpho)syntax, discourse, reading and spelling (Leonard, 2014; McArthur, Hogben, Edwards, Heath, & Mengler, 2000), and their language problems frequently co-occur with deficits in attention, working memory (Ebert & Kohnert, 2011; Montgomery, Evans, & Gilliam, 2018), and motor skills (Hill, 2001). Despite large heterogeneity in problems observed across children with DLD, almost all children with DLD exhibit problems with morphosyntax and phonological processing. Problems in these areas are therefore seen as clinical markers of the disorder (Leonard, 2014). Another criterion is that the problems observed in children with DLD cannot be attributed to neurological damage, hearing impairment, intellectual disability or unfavourable psycho-social/educational conditions.

One of the hypotheses on the origins of DLD states that the disorder may be the result of a statistical learning deficit (Evans et al., 2009; Hsu & Bishop, 2014a; Obeid et al., 2016; Wijnen, 2013). This hypothesis builds on two observations: the first one being that children with DLD often exhibit problems with linguistic aspects that require sensitivity to structural dependencies (e.g., English subject–verb agreement could also be described as a nonadjacent dependency between third person singular and verb-plus-[s] marking). The second observation concerns the comorbidity with problems in nonverbal areas. Following this, it has been argued that the linguistic problems observed in DLD may stem from nonverbal cognitive processing deficits that are thought to be related to language acquisition, amongst which is statistical learning (but see also various theories that claim for language-specific origins of the disorder: e.g., Grammatical Agreement Deficit Account, Clashen, 1989; Extended Optional Infinitive Account, Rice, Wexler, & Cleave, 1995).

The statistical learning deficit hypothesis is related to another general (non-language-specific) learning deficit hypothesis, namely the procedural learning deficit hypothesis (Ullman & Pierpont, 2005; Ullman, Earle, Walenksi, & Janacsek, 2020). The procedural learning deficit hypothesis states that the profile of language problems observed in children with DLD reflects a dissociation of children's functioning of the procedural learning system and their functioning of the declarative learning system. These two learning systems are not specific to language, but may be involved in learning rule-based aspects of language (procedural learning) and non-rule-based aspects of language (declarative learning). According to the procedural learning deficit hypothesis, children with DLD have a deficit in their procedural learning system that may explain their difficulties with sequenced aspects of language (e.g., grammar). By contrast, declarative learning is argued to be intact in this group of children and therefore, those language aspects that are learned via this system (e.g., word knowledge, irregular grammar) are relatively spared in children with DLD.

Different from the procedural learning deficit hypothesis, the statistical learning deficit hypothesis does not explicitly differentiate between procedural and declarative aspects of language learning. Statistical learning accounts assume that all aspects of language are learned via statistical computations. Different aspects of language may require different types statistical computations (sequential, distributional, cross-situational), though. Children with DLD may have difficulties with all these different types of computations, which may explain why the observed problems in this group of children range from problems with vocabulary building to establishing grammatical relations.

1.1.3 Measuring statistical learning

In the laboratory, participants' sensitivity to differences in sequential structure is commonly tested by exposing participants to structured stimuli and then measuring their knowledge of the structure. As an example, in the auditory verbal domain, participants can be exposed to an artificial language that may consist of a continuous stream of nonsense syllables from which words can be detected on the basis of differences in transitional probabilities between syllables within words and syllables across word boundaries (e.g., the string *bupadadutaba* consists of the two words: *bupada* and *dutaba;* Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). Alternatively, the artificial language may consist of a string of nonsense words that are defined by (non)adjacent dependencies between specific elements within a string. For instance, in the utterances *tep kasi lut, tep wadim lut, tep palti lut,* the first element (e.g., *tep*) and the third element (e.g., *lut*) form a nonadjacent dependency and thus the transitional probability between these two elements is 1 whereas the transition probability between the first and second word of the string (e.g., *tep* and *wadim*) and between the second word and third word of the string (e.g., *wadim* and *lut*) is lower (Gómez, 2002). Participants' sensitivity to the regularities can be measured via online measures of learning and offline measures of learning. Online measures of learning, such as response times, are taken during the exposure phase, whereas offline measures, such as grammaticality judgments, are taken after learning took place. As will become clear from the individual chapters in this dissertation, it is still a matter of debate, which criteria such measures should meet, and what constitutes the best online measure and the best offline measure of learning.

1.1.4 Outline and research questions

The present dissertation aims to provide an in-depth overview of statistical learning and its relation to language proficiency in children with and without DLD as such overviews are scarce (Erickson & Thiessen, 2015). The studies described in chapters 2 to 6 of this dissertation (see below) all focus on different aspects of statistical learning, language proficiency and the relation between these two. Eventually, a synthesis of all these studies (**Chapter 7**) addresses the three main aims of this dissertation, namely (1) whether we can detect statistical learning at the individual and group level, (2) whether individual differences in statistical learning ability correlate with language proficiency and (3) whether the problems observed in children with DLD can be explained by a statistical learning deficit that is independent of modality, domain and specific dependency type to be learned

Chapter 2 provides a quantitative overview (meta-analysis) of what is currently known on auditory verbal statistical learning in people with and without DLD, and it provides an estimate of the size of the auditory verbal statistical learning deficit in people with DLD. This chapter may also reveal whether concerns about the existence of a publication bias in the literature on statistical learning in atypically developing children are warranted (Schmalz et al., 2017; van Witteloostuijn et al., 2017).

Chapter 3 reports on the development of a novel, child-friendly online measure of statistical learning that can be used to assess auditory nonadjacent dependency learning in primary-school-aged children. This novel measure assesses the size of the "disruption peak" that occurs in children's response time pattern when a long stretch of stimuli with nonadjacent dependencies is disrupted by presenting stimuli without such dependencies.

Chapter 4 assesses whether children with DLD as compared to their typically developing peers have an auditory verbal nonadjacent dependency learning deficit (using the novel measure that is described in Chapter 3). The detection of such dependencies seems crucial in learning the morphosyntactic rules of a language. As almost all children with DLD have problems with morphosyntax, it is interesting to investigate their sensitivity to nonadjacent dependencies, as a specific type of statistical regularity. In investigating this issue, this chapter not only compares nonadjacent dependency learning between children with and without DLD, but also explores whether individual differences in nonadjacent dependency learning are associated with individual differences in grammatical proficiency.

Chapter 5 extends the focus from auditory verbal statistical learning to visual nonverbal statistical learning. The main aim of this chapter is to assess whether children with DLD have a visual nonverbal statistical learning deficit as compared to their typically developing peers. Assessing the presence and size of a visual statistical learning deficit is important if one wants to claim that children with DLD have a non-language-specific statistical learning deficit. Furthermore, as visual statistical learning has also been proposed to underlie literacy development in typically developing children (Arciuli & Simpson, 2012, von Koss Torkildsen, Arciuli, & Wie, 2019), this chapter also explores whether individual differences in visual statistical learning ability among children with DLD are associated with individual differences in literacy. The latter is interesting because children with DLD exhibit large individual differences in literacy performance: approximately half of the children with DLD have problems with reading and/or spelling in addition to their problems with oral language (McArthur et al., 2000).

By means of an experiment and meta-analysis **Chapter 6** addresses the association between serial reaction time task performance – a measure of nonverbal visuomotoric statistical learning – and grammatical proficiency in children with and without DLD. Three questions are addressed. First, we try to conceptually replicate the finding that children with DLD have a deficit in their detection of sequences in the nonverbal visuomotoric domain (experiment). Second, we assess the strength of the proposed correlation between children's nonverbal visuomotoric statistical learning and grammatical proficiency (experiment) and place this outcome in the context of previous work on this topic (meta-analysis). Third, it is explored whether the strength of the proposed correlation differs between children with and without DLD (meta-analysis).

1.1.5 Embedding of this dissertation and terminology used

The studies described in this dissertation are part of a larger project on the relation between statistical learning and grammar and literacy development in children (project "Progracy"). Progracy features two other projects, one on the relation between statistical learning and language proficiency in children with developmental dyslexia (studies conducted by Merel van Witteloostuijn) and the other on the developmental trajectory of statistical learning and its relation to language in typically developing children (studies led by Judith Rispens). Though reiterated in the relevant chapters, it may be good to mention that the same group of children that is described in **Chapter 3** of this dissertation also participated in the visual nonverbal statistical learning experiment described in van Witteloostuijn, Lammertink, Boersma, Wijnen and Rispens (2019). Also, as explained in each of the relevant chapters, there is overlap between the typically developing children that participated in the experiments described in this dissertation (**Chapters 4, 5** and **6**) and the typically developing children that participated in the studies described by Merel van Witteloostuijn (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019a, 2019b, submitted). Finally, within the scope of this dissertation, the results of the experiments described in **Chapters 4**, **5** and **6** come from the same children, therefore the sections describing the recruitment of these children and the sections describing their characteristics overlap.

It may also be good to provide an explanation for the inconsistency in the labeling of children with DLD throughout this dissertation. In 2015, when Progracy started, the most commonly used label to refer to children with a language disorder that is not associated with a known biomedical etiology was specific language impairment (but see Bishop, 2014 for alternative labels). Only short after publication of the first paper (Chapter 2) of this dissertation, Bishop, Snowling, Thompson, and Greenhalgh (2017) came with the recommendation to use the term developmental language disorder when referring to this group of children. Following this recommendation, the term DLD is used in all subsequent publications and chapters of this dissertation. The recommendation by Bishop et al. (2017) also had consequences for the inclusion of children in a clinical research sample. Formally, below average nonverbal intelligence and the co-occurrence of other neurodevelopmental language disorders would preclude the diagnosis of specific language impairment. With the new recommendations these two criteria no longer hold, so that children may be included in a clinical research sample of DLD while having below-average nonverbal intelligence and while having additional neurodevelopmental disorders. Importantly, at the moment that recruitment for the studies in this dissertation started, the Dutch criteria for diagnosing children with unexplained language difficulties as having DLD, did not follow these new recommendations (Stichting Simea, 2014). Therefore, all children with DLD that participated in the studies described in this dissertation have at least average nonverbal intelligence and have not been diagnosed with other neurodevelopmental language disorders.

Finally, all the data and scripts for analyses used for the studies described in this dissertation are openly available at Open Science Framework (OSF) project pages, and all publications that follow from this dissertation are open access. Therefore, we hope that this dissertation plays an exemplary role in making experimental research more transparent and available.

Chapter 2

Statistical learning in specific language impairment: A meta-analysis

This chapter is a slightly modified version of the paper that was published as:

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.

Data and scripts for analyses: https://osf.io/4exbz/

Abstract

The current meta-analysis provides a quantitative overview of published and unpublished studies on statistical learning in the auditory verbal domain in people with and without specific language impairment (SLI). The database used for the meta-analysis is accessible online and open to updates (Community-Augmented Meta-Analysis), which facilitates the accumulation and evaluation of previous and future studies on statistical learning in this domain. A systematic literature search identified 10 unique experiments examining auditory verbal statistical learning in 213 participants with SLI and 363 without SLI, aged between 6 and 19 years. Data from qualifying studies were extracted and converted to Hedges' *g* effect sizes. The overall standardized mean difference between participants with SLI and participants without SLI was 0.54, which was significantly different from 0 ($p < .001$, 95% confidence interval [0.36, 0.71]). Together, the results of our meta-analysis indicate a robust difference between people with SLI and people without SLI in their detection of statistical regularities in the auditory verbal input. The detection of statistical regularities is, on average, not as effective in people with SLI compared to people without SLI. The results of this meta-analysis are congruent with a statistical learning deficit hypothesis in SLI.

2.1 Introduction

Natural languages are structured at the level of sound (phonology), word formation (morphology), and sentence (syntax). These structures are reflected by statistical regularities in speakers' verbal output. Children learning their native language unconsciously detect and extract these regularities (Romberg & Saffran, 2010). This process, called statistical learning, is thought to be fundamental for the earliest stages of language acquisition (Evans, Saffran, & Robe-Torres, 2009). Two types of statistical learning are generally distinguished: distributional statistical learning and sequential statistical learning. Distributional statistical learning is about the detection of frequencies with which certain linguistic elements or structures occur. Sequential statistical learning concerns the detection of the sequential ordering and co-occurrence of concrete elements (e.g., syllables) in the auditory input in time (Kerkhoff, de Bree, & Wijnen, submitted). This meta-analysis focuses on sequential statistical learning, and therefore, from here onward, the term statistical learning refers to sequential but not distributional statistical learning.

Individual performance on statistical learning tasks has been shown to predict sentence comprehension (Misyak & Christiansen, 2012), processing of relative clause sentences with long distance dependencies, and lexical and oral language skills in participants' native language (Evans et al., 2009; Mainela-Arnold & Evans, 2014). Because tracking statistical patterns appears crucial for language acquisition and people differ in their ability to do this, it is not surprising that deficits in the ability to detect statistical patterns and relations in the input have been put forward as an explanation for impairments of language acquisition, notably specific language impairment (SLI; Evans et al., 2009; Hsu & Bishop, 2011; Ullman & Pierpont, 2005). A considerable number of studies looked at the domain specificity of this type of learning deficit in SLI. A recent meta-analysis by Obeid, Brooks, Powers, Gillespie-Lynch, and Lum (2016) summarized these findings and concluded that people with SLI perform worse on statistical learning tasks compared with typically developed people but that this difference in performance did not vary as a function of task modality (visual; visuomotoric and auditory) or age. The current meta-analysis provides a more extensive quantitative investigation of the difference in statistical learning ability between people with and without SLI[1] in the auditory domain. Different from Obeid et al. (2016), our focus is on the auditory verbal domain. Specifically, we were interested to see whether a difference in statistical learning performance between people with and without SLI varied as a function of linguistic level (word segmentation vs. grammar) or age at which learning took place.

2.1.1 Statistical learning in de laboratory

Many experimental studies of statistical learning focus on learning dependencies. These dependencies can be learned at different linguistic levels (e.g., word segmentation vs. grammar). We first discuss examples of artificial word segmentation studies followed by examples of artificial grammar learning studies.

In experiments that simulate word segmentation, participants are exposed to a continuous stream of syllables that are organized according to a set of statistical regularities. The stimuli are designed in such a way that transitional probabilities of sequences of certain adjacent syllables are higher than transitional probabilities of other adjacent syllables (continuous relationship), reflecting word boundaries (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). After exposure, participants perform a lexical decision task (or word recognition task via a preferential looking paradigm in the case of infant studies) in which they hear sequences of syllables that had high transitional probabilities in the exposure phase (reflecting words) as well as sequences of syllables that had low to zero transitional probabilities in the exposure phase. Accordingly, adult participants have to indicate whether the words they are presented with are part of the language they were familiarized with or not. In infant studies, the listening times to the sequences of syllables with high transitional probabilities versus sequences of syllables with low transitional probabilities are compared. Results show that adults and infants are able to distinguish such artificial high probability words from artificial low probability words on the basis of adjacent transitional probabilities.

Contrary to word segmentation studies, the stimuli in artificial grammar learning studies consist of already segmented words that have primary stress and

[1] When we speak of people without SLI, we mean people who are matched in age and/or intelligence to participants with SLI (see Table 2.1), who have no reported (history of) hearing, language, or learning problems and no reported (history of) neurological impairment or illness.

minimal coarticulation and are separated by pauses. Artificial grammar learning studies aim to resemble grammatical phenomena present in natural language. In natural language, for example, grammatical relations are present among functional elements (e.g., *is* and *ing*) across interleaved lexical elements (e.g., *Grandma is singing*; example taken from Sandoval & Gómez, 2013). In experimental designs that test this type of learning, participants are exposed to strings generated, unknown to the learner, by a miniature artificial grammar. The grammar follows a set of nonadjacent (discontinuous) dependency relations (Gómez, 2002), a set of predictive relations (cf. Saffran, 2002), or a set of finite rules (finite state grammar; Gómez & Gerken, 1999). The procedure of artificial grammar learning designs is similar to the procedure in word segmentation studies: After a period of exposure to the language, participants are tested with strings that either conform to the grammar (grammatical items) or that violate the grammar (ungrammatical items), and participants have to indicate whether the string they hear is grammatical or ungrammatical. More important, participants are asked to judge strings with elements that they have heard during the familiarization phase of the experiment as well as strings with novel elements that they have not heard before to test for generalization of the rule (although not all artificial grammar learning studies test for generalization; see Grama, Kerkhoff, & Wijnen, 2016).

2.1.2 Cognitive processes involved in auditory verbal statistical learning

As stated in our operational definition, statistical learning requires sensitivity to regularities in the input (e.g., statistical cues like transitional probabilities in word segmentation and [non]adjacent dependencies in artificial grammar learning). However, there are also other cognitive processes involved in auditory verbal statistical learning such as phonological awareness (the ability to analyse and manipulate incoming phonemes and syllables), verbal short-term memory, and verbal working memory. Both word segmentation and artificial grammar learning involve the temporary storage of incoming input, which is necessary to pick up the statistical regularities between elements in the input (verbal short-term memory). In addition, artificial grammar learning, compared with word segmentation, requires processing of long distance dependencies and generalizing those dependencies to novel items. Long distance dependencies have been argued to put more demand on working memory than adjacent dependencies (see, e.g., theoretical models on resource limitation of Gibson [1998]), and generalization is more demanding than recognition of items previously introduced (Thompson & Newport, 2007). Therefore, we hypothesize that artificial grammar learning, compared with word segmentation, is more demanding on working memory capacity. In the following section, we discuss how this difference between both levels of learning might disadvantage individuals with SLI in their auditory verbal statistical learning performance.

2.1.3 Statistical learning in SLI

In natural language, SLI is characterized by problems at the grammatical level (e.g., subject–verb agreement, past–tense marking; Leonard, 2014) as well as at the word segmentation level (e.g., lexical–phonological deficits observed in gating and nonword repetition tasks; see Mainela- Arnold, Evans, & Coady, 2010; Graf Estes, Evans, & Else-Quest, 2007). In artificial language, we see a similar pattern: Most studies investigating auditory verbal statistical learning in SLI show that participants without SLI outperform participants with SLI both in word segmentation and in grammar learning tasks (word segmentation: Evans et al., 2009; grammar: Hsu, Tomblin, & Christiansen, 2014; Lukács & Kemény, 2014; Mainela-Arnold & Evans, 2014; Mayor-Dubois, Zesiger, van der Linden, & Roulet-Perez, 2014). It is known that people with SLI exhibit deficits in verbal short-term memory and verbal working memory as well (Archibald & Gathercole, 2006; Marton, Eichorn, Campanelli, & Zakariás, 2016; Montgomery, 2003). As these processes are involved in auditory statistical learning, it might well be the case that these deficits influence the auditory verbal statistical learning abilities of people with SLI. Previous research, however, suggests that memory problems cannot solely explain auditory statistical learning problems. For example, individuals with SLI have problems with statistical learning in the nonverbal domain (Lum, Conti-Ramsden, Morgan, & Ullman, 2014; Lum, Conti-Ramsden, Page, & Ullman, 2012; Obeid et al., 2016), which are unlikely to be caused by verbal short-term and working memory problems. In addition, Hsu and Bishop (2014a) report poor verbal sequence learning in children with SLI, even after controlling for limitations of verbal short-term memory (Hebb repetition task). Taken together, results of previous studies are congruent with the hypothesis that SLI is associated with a “statistical learning disadvantage.” The magnitude and moderators of this disadvantage, however, are unknown. Therefore, the primary purpose of the current meta-analysis was to assess the magnitude of this statistical learning disadvantage in the auditory verbal domain. The second goal was to explore the potential impact of linguistic level and age at which learning takes place. We wanted to explore whether the statistical learning disadvantage is more severe in artificial grammar learning than word segmentation studies, as the former type of learning is more demanding on verbal working memory capacity, which is generally affected in SLI. With the second meta-regression, we explore whether age moderates the statistical learning disadvantage. Previous studies investigating the influence of age in statistical learning have provided mixed results. Obeid and colleagues (2016) reported no effect of age on statistical learning differences between people with and without SLI across different modalities of learning. Lum and colleagues (2014), however, reported smaller differences in visuospatial statistical learning performance between people with and without SLI for older compared with younger participants. Likewise, studies investigating the developmental trajectory of statistical learning in typically developing people have reported mixed results. Some studies report that there is no evidence for a difference in statistical learning performance between adults and children (visual domain: Kirkham, Slemmer, & Johnson, 2002; auditory domain: Saffran et al., 1997), whereas others do report that statistical learning performance improves with age (visual domain: Arciuli & Simpson, 2011; auditory domain: Lukács & Kemény, 2015; visuospatial: Meulemans, van der Linden, & Perruchet, 1998).

2.1.4 The present study

The current meta-analysis provides an estimate of the magnitude of the statistical learning disadvantage in people with SLI by means of a quantitative overview of both published and unpublished studies that investigate statistical learning in the auditory verbal domain in people with and without SLI. In a first step, we calculated the standardized averaged mean difference (effect size measure) in performance on statistical learning tasks in people with and without SLI. In a second analysis, we explored whether the effect size measure was moderated by linguistic level (word segmentation vs. grammar) and age.

2.2 Method

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement to organize the current meta-analysis (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). Effect size calculations were done in the statistical software R (R Core Team, 2017). Formulas were implemented via the R *compute.es* package (Del Re, 2013), and statistical analyses on the effect size measures were conducted with the R *meta* (Schwarzer, 2015) and *metafo*r (Viechtbauer, 2010) packages.

2.2.1 Literature search

Systematic searches for empirical articles were conducted in February 2016 using a combination of prespecified key word combinations (details of all key words, Boolean operators, and syntax used for each database can be found in Appendix A2.1). We conducted our searches in five different sources including PubMed, Education Resources Information Center, PsycINFO, Linguistics and Language Behavior Abstracts, and Open Access Theses and Dissertations. In addition, we asked experts in the field to inform us of any published or unpublished studies via two different calls (LINGUIST List and Cogdevsoc list; July 2016). These combined searches yielded 161 articles (PubMed: 26 hits, Education Resources Information Center: 25 hits, PsycINFO: 64 hits, Linguistics and Language Behavior Abstracts: 38 hits, Open Access Theses and Dissertations: five hits, and experts in the field: three hits).

2.2.2 Inclusion criteria and study selection

To be included in the meta-analysis, studies were required to meet the following criteria: (a) A study should report on original empirical research data. Both published and unpublished studies were eligible, including articles in refereed journals, nonrefereed journals, dissertations, and conference presentations; (b) a study should have an experimental design that tests sequential statistical learning in the auditory verbal domain assessed via a word segmentation, grammaticality judgment, or related task; (c) as we aimed to test whether participants implicitly detected the statistical regularity, participants should not receive any explicit instruction or feedback regarding the underlying structure of the artificial language to be learned or on their behavior during the training or test phase; and (d) selected studies include one group of participants with SLI and one group of age-matched controls who do not have language impairments. More important, we only included studies that identified participants with SLI on the basis of inclusion and exclusion criteria typical for SLI. Therefore, studies had to report scores on standardized language tests[2] or use a test battery that differentiates

[2] Participants with SLI scored at least 1.25 standard deviations below age norms.

between participants with and without a history of SLI (e.g., Tomblin battery; Tomblin, Freese, & Records, 1992; see Table 2.1). In addition, a nonverbal intelligence measure[3] and no history of neurological or emotional delays should be reported for both participant groups. It is important to mention that the inclusion and exclusion criteria for SLI vary across the studies in our sample (see Table 2.1). We only included studies, however, that based their inclusion criteria on both standardized language tests and intelligence scores. If studies failed to report on one of these criteria (or if information on these criteria could not be confirmed via contact with the authors), the study was excluded from the analysis. In addition, when studies included children with nonverbal intelligence below 80, the control group and the group with SLI had to be matched on nonverbal intelligence to ensure that differences in statistical learning performance are not the result of lower intelligence scores. Finally, to be included in the analysis for the current article, studies had to be conducted before September 2016 (but see footnote 4). However, as our database is accessible online and open to update, future studies can be added, which facilitates accumulation and evaluation of previous and future studies on statistical learning in this domain (Tsuji, Bergmann, & Cristia, 2014). No start date for publications was set to find as many studies as possible. For an overview of the exact inclusion and exclusion criteria for the studies in our final sample, see Table 2.1.

After removing duplicates, 81 studies (78 published articles and three unpublished conference posters) remained. Two reviewers independently conducted the study selection procedure. In a first step, both reviewers performed a full-text inspection of the 19 studies (16 published articles and three nonpublished conference posters) that were selected, based on screening of the title and abstract. The reviewers independently screened these full-text articles and posters according to the inclusion criteria. There was 95% (18/19 studies) agreement on the selection of these full-text studies (eight studies included, 10 studies excluded, one study for discussion). After discussion, the reviewers decided not to include the one study they had disagreed on because participants in this study had received feedback on their behavior during the test phase (von Koss Torkildsen, Dailey, Aguilar, Gómez, & Plante, 2013). As a result, the initial

[3] Nonverbal intelligence had to fall within the normal range (>80), or when the lower limit of intelligence was <80, the control group and the group with SLI had to be matched on nonverbal intelligence.

final selection consisted of eight studies (five published articles and three nonpublished conference posters).[4] For a visual representation of the literature search procedure, see Figure 2.1.

Four of the eight studies reported multiple individual experiments or multiple outcomes per participant group (Evans et al., 2009; Grunow, Spaulding, Gómez, & Plante, 2006; Hsu & Bishop, 2014a; von Koss Torkildsen, 2010). If the data necessary to compute the individual effect size were available for each experiment separately and the groups of participants tested in the experiments were independent (i.e., different participants), all of the experiments of that study were included in the meta-analysis. Only the study of Hsu et al. (2014) met these criteria. For the other three studies with multiple experiments (Evans et al., 2009; Grunow et al., 2006; von Koss Torkildsen, 2010), only one effect size measure was incorporated into the final analysis (for more details on our decisions with respect to this part, see the subsection Effect size calculation). This resulted in a final sample of 10 experiments.

2.2.3 Sample description

The eight studies (10 experiments) we included in our analysis were published (six studies) or presented (two studies) between 2006 and 2017 (see footnote 4). The experiments collectively examined 213 participants with SLI and 363 controls, all between 6 and 19 years old. The dependent variable was slightly different across the 10 experiments. In six experiments, the outcome variable was the overall accuracy score on a grammaticality judgment task; in three experiments, the outcome variable was the overall accuracy score on a word segmentation task; and in one experiment, the outcome variable was an event-related potential (ERP: P600).

2.2.4 Effect size calculation

For each individual experiment, we calculated the effect size (Hedges' *g*) as the standardized mean difference (SMD)[5] in performance between the participants

[4] During the review of our current meta-analysis, the poster of Haebig and colleagues got published as an article in *The Journal of Child Psychology and Psychiatry*. Therefore, the final data set consists of six published articles and two nonpublished conference posters.

[5] The standardized mean difference expresses the size of the effect in each study relative to the variability observed in that study (Higgins & Green, 2011).

with and without SLI. The SMD was chosen over the raw mean difference, because the dependent variables differed across studies (ERP amplitude vs. accuracy scores).

All formulas used to calculate the SMD and the approximation of the variance of the SMD for each individual experiment are shown in Appendix A2.2 and were taken from the R *compute.es* package (Del Re, 2013). The effect size was calculated so that positive values indicated that the participants without SLI outperformed the participants with SLI. For seven of the 10 experiments (Evans et al., 2009; Evans, Hughes, Hughes, Jackson, & Fink, 2010; Haebig, Saffran, & Weismer, 2017; Hsu et al., 2014; Lukács & Kemény, 2014), the SMD was calculated with the mean overall accuracy scores and the standard deviation scores for both participant groups (*mes2* function in the R *compute.es* package). For two experiments (Grunow et al., 2006; von Koss Torkildsen, 2010), the SMD was calculated from the reported *F* statistic on the main effect of group (*fes* function from the *compute.es* package), and for one experiment (Mayor-Dubois et al., 2014), the reported *t* statistic was used to calculate the SMD (*tes* function in the *compute.es* package).

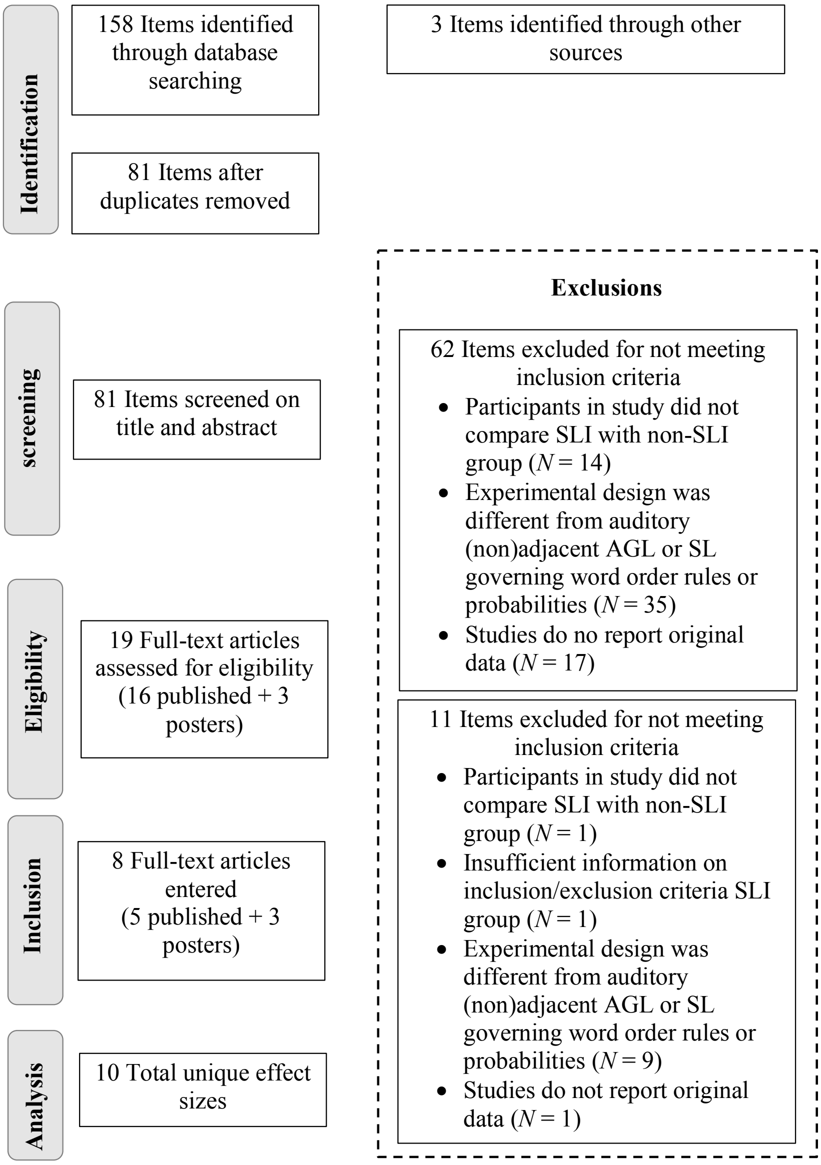

Identification
158 Items identified through database searching
3 Items identified through other sources
81 Items after duplicates removed
Exclusions
screening
81 Items screened on title and abstract
62 Items excluded for not meeting inclusion criteria
• Participants in study did not compare SLI with non-SLI group (N = 14)
• Experimental design was different from auditory (non)adjacent AGL or SL governing word order rules or probabilities (N = 35)
• Studies do no report original data (N = 17)
Eligibility
19 Full-text articles assessed for eligibility (16 published + 3 posters)
11 Items excluded for not meeting inclusion criteria
• Participants in study did not compare SLI with non-SLI group (N = 1)
• Insufficient information on inclusion/exclusion criteria SLI group (N = 1)
• Experimental design was different from auditory (non)adjacent AGL or SL governing word order rules or probabilities (N = 9)
• Studies do not report original data (N = 1)
Inclusion
8 Full-text articles entered (5 published + 3 posters)
Analysis
10 Total unique effect sizes

Figure 2.1 Flowchart indicating data exclusion at each stage of the literature search procedure. AGL = artificial grammar learning; SL = statistical learning.

Table 2.1 Overview of the study sample characteristics for each individual experiment included in our meta-analysis

Study	Native Language	Sample size		Mean age (y;m)		SLI inclusion criteria	Matching participants within study
		SLI	TD	SLI	TD		
Evans et al. (2009)	English	35	78	9;6	7;9	(a) Nonverbal IQ >85 (LIPS-R) (b) Normal hearing (c) Normal corrected vision (d) Normal oral speech and motor abilities (e) Expressive language composite score >1.5 *SD*s below mean (CELF-3)	Age, nonverbal IQ
Evans et al. (2010)	English	14	14	16;5	15;6	(a) Nonverbal IQ >80 (LIPS-R, WISC-R, WPPSI) (b) Normal hearing (c) No major neurological abnormalities (d) Absence of other dev. disorders (e) Expressive language composite score >1.5 *SD*s below mean (CELF-R) (f) Receiving speech and language services	Age

(*Table continues*)

Table 2.1 (*Continued*)

Study	**Native Language**	**Sample size**		**Mean age (y;m)**		**SLI inclusion criteria**	**Matching participants within study**
		SLI	**TD**	**SLI**	**TD**		
Grunow et al. (2006)	English	22	22	19;1	18;5	(a) Normal nonverbal IQ (Test of nonverbal intelligence III[a]) and IQ should not differ from TD (b) Normal hearing (c) No history of seizures/head trauma (d) No diagnosis of ADHD (e) Language impairment status attested via Tomblin battery[b]	Age, nonverbal IQ
Haebig et al. (2017)	English	25	30	10;4	10;4	(a) Normal nonverbal IQ (WISC-4) (b) Language assessment via PPVT-4 and CELF-4	Age, Nonverbal IQ

(*Table continues*)

Table 2.1 (*Continued*)

Study	Native Language	Sample size		Mean age (y;m)		SLI inclusion criteria	Matching participants within study
		SLI	TD	SLI	TD		
Hsu et al. (2014)[c]	English	20	20	13;8	14;3	(a) Nonverbal IQ should not differ from controls (WISC-3) (b) Significant poorer language scores on composite scores from CELF-3, PPVT-R, CREVT, and listening QRI-2	Age, Nonverbal IQ
Hsu et al. (2014)[d]	English	20	20	14;1	14;2	(a) Nonverbal IQ should not differ from controls (WISC-3) (b) Significant poorer language scores on composite scores from CELF-3, PPVT-R, CREVT, and listening QRI-2	Age, Nonverbal IQ
Hsu et al. (2014)[e]	English	20	20	14;2	13;9	(a) Nonverbal IQ should not differ from controls (WISC-3) (b) Significant poorer language scores on composite scores from CELF-3, PPVT-R, CREVT, and listening QRI-2	Age, Nonverbal IQ

(*Table continues*)

Table 2.1 (*Continued*)

Study	Native Language	Sample size		Mean age (y;m)		SLI inclusion criteria	Matching participants within study
		SLI	TD	SLI	TD		
Lukács & Kemény (2014)	Hungarian	29	87	9;1	9;1	(a) Nonverbal IQ >85 (RAVEN) (b) Normal hearing (c) No neurological impairment (d) 1.5 *SD*s below age norms at two of the language tests assessed (Hungarian PPVT, TROG, MAMUT, NWR)	Age
Mayor-Dubois et al. (2014)	French	18	65	10;1	10;0	(a) Nonverbal IQ >80 (WISC-4) (b) SLI diagnosis confirmed via standardized language assessment by speech and language therapists (c) Participants with SLI were still pursuing speech and language therapy	Age

(*Table continues*)

Table 2.1 (*Continued*)

Study	Native Language	Sample size		Mean age (y;m)		SLI inclusion criteria	Matching participants within study
		SLI	**TD**	**SLI**	**TD**		
von Koss Torkildsen (2010)	Norwegian	14	14	6;1	5;9	(a) Nonverbal IQ >70 (b) Normal hearing and (corrected) vision (c) No history of epilepsy, cerebral palsy or brain haemorrhage (d) No structural abnormalities in speech system (assessed by speech pathologist) (e) 1.25 *SD*s below age norms at standardized language tests (assessment done by pedagogical-psychological services)	Age, Nonverbal IQ

Note. y = years; m = months; TD = typically developing; IQ = intelligence; LIPS-R = Leiter International Performance Scale–Revised (Roid & Miller, 1997); CELF-3 = Clinical Evaluation of Language Fundamentals–Third Edition (Semel et al., 1995); WISC-R = Wechsler Intelligence Scale for Children–Revised (Wechsler, 1974); WPPSI = Wechsler Preschool and Primary Scale of Intelligence (no version reported); CELF-R = Clinical Evaluation of Language Fundamentals–Revised (Semel et al., 1987); WISC-4 = Wechsler Intelligence Scale for Children–Fourth Edition (Wechsler, 2003); PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007); CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition (Semel et al., 2003); WISC-3 = Wechsler Intelligence Scale for Children–Third Edition (Wechsler, 1991); PPVT-R = Peabody Picture Vocabulary Test–Revised (Dunn & Dunn, 1981); CREVT = Comprehensive Receptive and Expressive Vocabulary Test: Adult (Wallace & Hammill, 1997); QRI-2 = Qualitative Reading Inventory–Second Edition (Leslie & Caldwell, 1995); RAVEN = Raven Progressive Matrices and Raven Coloured Matrices (Raven et al., 1987); PPVT = Peabody Picture Vocabulary Test (Csányi, 1974); TROG = Test for Reception of Grammar (Lukács et al., 2012); MAMUT = Magyar Mondatutánmondási Teszt (Hungarian Sentence Repetition Test; Kas & Lukács, 2011); NWR = nonword repetition task (Racsmány et al., 2005). [a]Brown et al. (1997). [b]Tomblin et al. (1992). [c]Participant characteristics for set size x = 24. [d]Participant characteristics for set size x = 12. [e] Participant characteristics for set size x = 2.

As mentioned in the Inclusion criteria and study selection section, it was not always possible to calculate multiple effect sizes for studies that ran multiple experiments. In the case of the Grunow et al. (2006) experiments, we calculated one effect size because the statistical information necessary to calculate a separate effect size for each of the different experimental conditions (low vs. high intervening X-element, generalization vs. nongeneralization items) was not available. Likewise, one effect size was obtained from the study by Evans at al. (2009), which reported on two different experiments. The second experiment was conducted 6 months after the first. The participants of Experiment 2, however, had all participated in Experiment 1, rendering the data of the second experiment correlated with a part of the data of the first experiment. A combined effect size, taking the correlation term between Experiments 1 and 2 into account, would have been the ideal solution because it would take into account the increased precision of within-subject measures (Borenstein, Hedges, Higgins, & Rothstein, 2009, pp. 28–30). However, it was impossible to determine the correlation term between the two experiments, because only parts of the data were correlated. Therefore, we included only the first experiment, which had twice as many participants as the second experiment. Last, von Koss Torkildsen (2010) recorded ERPs during both the exposure phase and the test phase. As we have no measures of performance during the exposure phase for the other studies in our sample, only the effect size measure of the ERPs recorded during the test phase is included.

Finally, we applied Hedges' *g* correction for small sample sizes to all 10 effect sizes, because most of the experiments had a sample size of less than 20 (Borenstein et al., 2009, p. 27).

2.3 Results

2.3.1 Publication bias

Meta-analyses are generally sensitive to publication bias. Publication bias reflects the tendency of a higher publication rate for studies with significant results compared with studies with nonsignificant results (Dickersin, 2005). Because it is more likely that published studies end up in a meta-analysis, the overall combined effect size might be overestimated when there is a publication bias in the sample used to compute the combined effect sizes (Borenstein et al., 2009, p. 278). In the current meta-analysis, we analysed funnel plot asymmetry as a potential indicator of publication bias (Egger, Smith, Schneider, & Minder, 1997).

In our funnel plot (see Figure 2.2), the effect size of a particular experiment is plotted against the standard error of that particular experiment. The standard error can be interpreted as a measure of experiment size, as generally experiments with fewer participants have higher standard errors. In the absence of publication bias, a funnel plot is symmetric and funnel shaped; large experiments appear toward the top (low standard error) of the plot and generally cluster around the mean effect size, whereas smaller experiments appear toward the bottom (higher standard error) of the graph and tend to be spread across a broader range of values. Visual inspection of our funnel plot (see Figure 2.2) seems to suggest asymmetry such that smaller experiments tend to have greater effect sizes (i.e., they appear more to the right side of the mean effect size than the left side). The latter could indicate publication bias, as small experiments are more likely to be found (or published) when the effect size is large compared with when the effect size is small. We performed a linear regression on funnel plot asymmetry (Egger et al., 1997). The test on funnel plot asymmetry was performed using the *regtest* function in the *metafor* (Viechtbauer, 2010) R package. The regression on funnel plot asymmetry was not significant (z = 1.52, p = .13). Therefore, we have no statistical evidence for a publication bias in the current sample.

2.3.2 Primary analysis: Effect size and heterogeneity

We estimated the average weighted SMD and heterogeneity of the sample with a random-effects model with the restricted maximum-likelihood estimator for the amount of heterogeneity. All 10 observed effect sizes and their weights were included to estimate the median effect size. No further moderator variables were specified in the model. Sample heterogeneity was assessed via Cochran's *Q* test for heterogeneity.

The overall weighted mean effect size and the observed effect sizes for the individual experiments are shown in Figure 2.3. The average observed weighted mean effect size (intercept) under our random-effects model (random effect = study) was 0.54 (SE = 0.09, 95% confidence interval [CI] [0.36, 0.70]). The observed effect size was significantly different from zero (z = 5.98, p = $2.2 \cdot 10^{-9}$) and positive, which indicates that people without SLI, on average, outperform people with SLI on statistical learning tasks in the auditory verbal domain. In other words, the value of 0.54 can be regarded as our estimate for the statistical learning disadvantage in people with SLI. Furthermore, the CI ranges from 0.36 to 0.70, indicating that we reliably detected any effect size up to 0.36, which means that we can speak of a moderate-to-large statistical learning disadvantage in people with SLI.

As a measure of heterogeneity, the total amount of variance between the experiments was $\tau^2 = 0.0$ (SE = 0.036). Cochran's *Q* test for heterogeneity was not significant ($Q(9) = 10.11$, $p = .34$). This means that there is no statistical evidence that the true effect sizes differ between the studies in our sample. It is important to note, however, that, whereas a significant *Q* test provides evidence that the true effects vary, a nonsignificant *Q* test alone should not be taken as evidence that the true effect sizes are consistent. The low number of experiments in our design could well explain the finding of nonsignificant heterogeneity (Borenstein et al., 2009, p. 113).

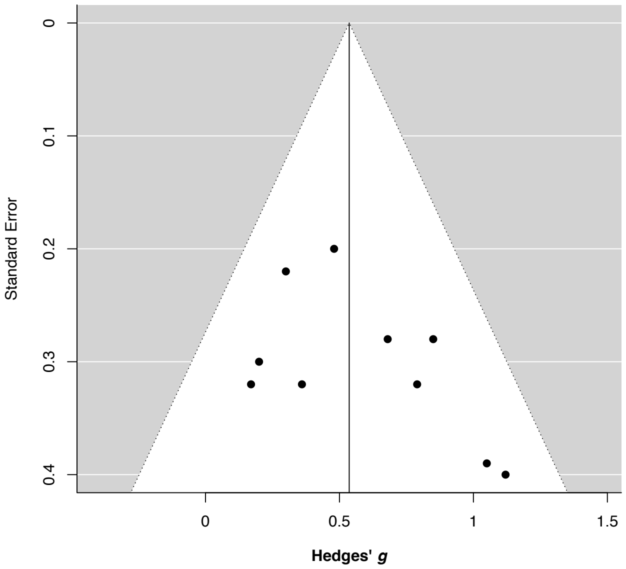


Figure 2.2 Funnel plot showing standard error of the effect size Hedges' *g* as a function of effect size. The vertical line indicates the overall model estimate. The triangle-shaped unshaded region represents a pseudo confidence interval region with bounds equal to ± 1.96 SE.

2.3.3 Secondary analysis: Meta-regressions on linguistic level and age

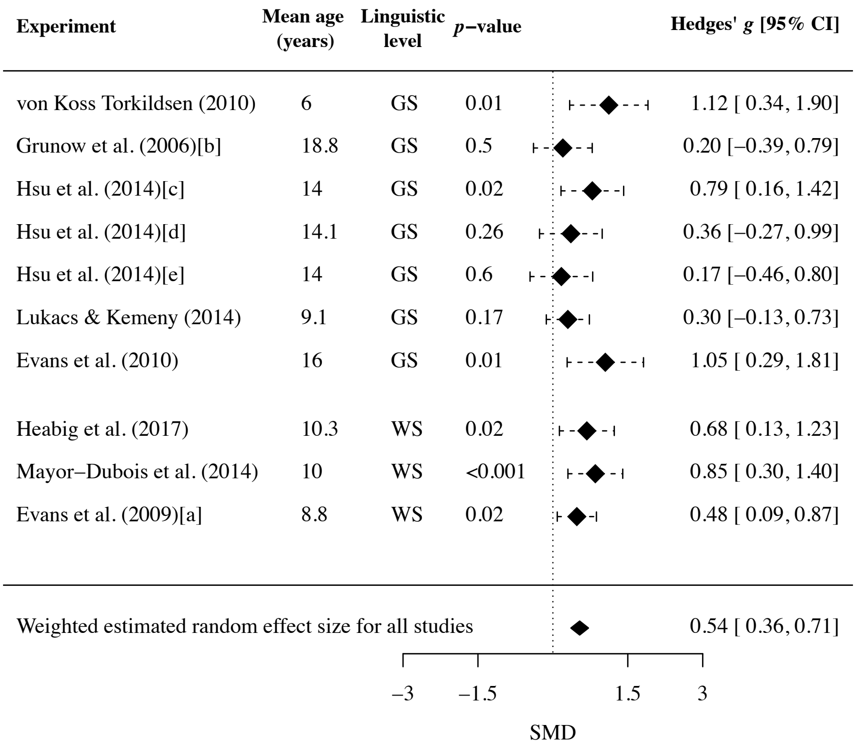
As mentioned in the Introduction, we were interested in seeing whether the linguistic level (word segmentation vs. grammar) and age at which the experiments were performed influence the SMD. We do realize, however, that our sample includes only 10 studies, which renders it unlikely that we will find a significant effect. Nevertheless, we decided to continue our meta-regression, as assessing the impact of the moderator variables linguistic level and age was part of our research question. As our moderator variables are correlated, the impact of both moderators is evaluated by means of two separate meta-regression models.

To assess whether the linguistic level at which the experiments were performed (word segmentation vs. grammatical structure) influences the SMD, we added linguistic level as a between-experiments moderator variable to the random-effects model described above. When we coded experiments at the word segmentation level as $-\frac{1}{2}$ and experiments at the grammatical level as $+\frac{1}{2}$, the resulting mixed-effects model detected no significant effect of linguistic level (estimate of the SMD difference = −0.15, SE = 0.18, $z = -0.80$, $p = .43$, 95% CI [−0.51, +0.21]).

As can be seen in Figure 2.3 (and Table 2.1), the studies in our sample included participants between 6 and 19 years old. To test for age effects, we fit a second meta-regression model with age in years (log-transformed) as the continuous predictor variable. The mixed-effects model detected no significant effect of age (estimate of the SMD difference = −0.10, SE = 0.11, $z = -0.91$, $p = .36$, 95% CI [−0.32, +0.12]).

In summary, we found no evidence that linguistic level or age influences the statistical learning disadvantage in people with SLI.[6] The potential effects of these moderators might be too small to detect with meta-regression due to the relatively small number of studies in our sample.

[6] In addition, we conducted an exploratory meta-regression with the moderator variable adjacency type. This regression revealed no significant effects either. As one of our reviewers pointed out, however, a meta-regression with the moderator variable adjacency type is problematic, as adjacency type is highly correlated with linguistic level (i.e., all word segmentation studies feature an adjacent dependency learning paradigm, whereas the artificial grammar learning studies featured a mix of adjacent and nonadjacent dependency types).



Experiment	Mean age (years)	Linguistic level	*p*–value	Hedges' *g* [95% CI]
von Koss Torkildsen (2010)	6	GS	0.01	1.12 [0.34, 1.90]
Grunow et al. (2006)[b]	18.8	GS	0.5	0.20 [–0.39, 0.79]
Hsu et al. (2014)[c]	14	GS	0.02	0.79 [0.16, 1.42]
Hsu et al. (2014)[d]	14.1	GS	0.26	0.36 [–0.27, 0.99]
Hsu et al. (2014)[e]	14	GS	0.6	0.17 [–0.46, 0.80]
Lukacs & Kemeny (2014)	9.1	GS	0.17	0.30 [–0.13, 0.73]
Evans et al. (2010)	16	GS	0.01	1.05 [0.29, 1.81]
Heabig et al. (2017)	10.3	WS	0.02	0.68 [0.13, 1.23]
Mayor–Dubois et al. (2014)	10	WS	<0.001	0.85 [0.30, 1.40]
Evans et al. (2009)[a]	8.8	WS	0.02	0.48 [0.09, 0.87]
Weighted estimated random effect size for all studies				0.54 [0.36, 0.71]

Figure 2.3 Forest plot showing overall and individual average weighted effect sizes (Hedges' *g*) and 95% confidence interval (CI). A positive effect size indicates that the control group outperformed the group with specific language impairment. GS = grammatical structures; WS = word segmentation; SMD = standardized mean difference. [a] = effect size over experiment 1a; [b] = combined effect size over both types of generalization and set size; [c] = effect size for set size x = 24; [d] effect size for set size x = 12; [e] effect size for set size x = 2.

2.4 Discussion

The primary purpose of our meta-analysis was to provide a quantitative overview of published and unpublished studies on auditory verbal statistical learning in SLI to evaluate the magnitude of the auditory verbal statistical learning disadvantage in people with SLI. We found that, on average, the detection of statistical regularities in the input was not as effective in people with SLI as in people without SLI (statistical learning disadvantage) and that this difference in performance was moderate to large. The results supplement the findings of Obeid et al. (2016) on statistical learning across different modalities in people with SLI.

Different from Obeid and colleagues, our focus was on statistical learning in the auditory verbal domain, which allowed us (a) to add five additional studies on statistical learning in this domain that were not included in the Obeid et al. study and (b) to further explore whether differences in statistical learning ability between people with and without SLI arise as a function of linguistic level. Following on the latter, the second goal of our meta-analysis was to investigate whether the magnitude of statistical learning disadvantage in people with SLI was moderated by the linguistic level (word segmentation vs. grammar) or age at which learning takes place. We did not find evidence that the difference in statistical learning performance between people with and without SLI is moderated by the linguistic level and age at which learning takes place. Although the absence of the effect of linguistic level is a null effect and therefore difficult to interpret, it is in line with previous research reporting absences of associations between verbal working memory and sequence repetition learning (Hsu & Bishop, 2014a; Lum et al., 2012). Alternatively, the potential influence of both moderators might have been too small to detect with our meta-regressions due to the relatively small number of studies in our sample.

In all, our results extend previous findings on a visual statistical learning disadvantage in SLI (Lum et al., 2012, 2014; Obeid et al., 2016) to the auditory verbal domain and underline the assumption of a general cognitive deficit in the implicit detection of statistical regularities and/or dependencies in people with SLI that contributes to the language problems seen in this population (see also Evans et al., 2009; Hsu & Bishop, 2011, 2014a; Ullman & Pierpont, 2005).

2.4.1 Relevance for clinicians working with SLI

The current meta-analysis provides evidence that people with SLI have more difficulties with statistical learning than people without SLI. These findings support the use of evidence-based interventions that facilitate and stimulate the detection of (statistical) regularities in the input for people with SLI. A concrete example of such a statistical learning–based intervention is the conversation recast treatment for morpheme errors in children with SLI (Plante et al., 2014). Plante and colleagues base their training method on findings from artificial language studies. In such studies, strings have an a X b structure in which the a and b elements always co-occur (Gómez, 2002). It has been found that participants only learn the nonadjacent dependencies when the variability (i.e., different numbers of X-elements) of the intervening X-element is high enough

(Gómez, 2002). Likewise, Plante et al. (2014) showed that children's use of trained morphemes improved for children who were trained on these morphemes in a high-variability context (24 different verbs). They found no evidence of such a treatment effect for children in the low-variability (12 different verbs) context. It thus seems that both people with and without language impairment benefit from variability and not only repetition in their language input (Plante et al., 2014). High variability facilitates rule learning rather than rote learning, as participants need to look for regularities and patterns in the input as soon as they notice that memorization is not an option in case of high variability (exceeding working memory capacity). These results suggest that clinicians working with children with SLI need to provide a great number of examples when explaining new rules.

2.4.2 Publication bias

We would like to stress that, although the regression on funnel plot asymmetry did not reach significance, one should always be cautious for the possibility of publication bias in the literature on auditory statistical learning in SLI. Such a potential publication bias relates to the validity of the classical statistical learning paradigms to measure statistical learning efficiency. Recently, more and more researchers stress the importance of an online measure of statistical learning (e.g., Bogaerts, Franco, Favre, & Rey, 2016; Isbilen, McCauley, Kidd, & Christiansen, 2017; Misyak, Christiansen, & Tomblin, 2010) or a test phase that is more sensitive to individual variation. As mentioned by Siegelman, Bogaerts, and Frost (2017), a large proportion of the participants in a statistical learning study perform at chance level. On the group level, test performance is usually just above chance, and an accuracy score higher than 60% is rarely obtained. For these reasons, we consider it likely that more research groups have unpublished (pilot) data on auditory statistical learning in SLI that did not yield statistical significance. Inclusion of these unpublished data could have made our estimates more precise, and we therefore invite researchers who have such unpublished null results to contribute to our Community-Augmented Meta-Analysis via https://osf.io/4exbz/.

2.4.3 Recommendations for future studies

The results of the current meta-analysis show that there is a moderate-to-large statistical learning disadvantage in people with SLI. The moderators of this disadvantage, however, remain unknown. Therefore, we recommend that future studies test the effects of potential moderators such as linguistic level and age within a single study in which the variables are within-subject predictors. Longitudinal designs can be used to test statistical learning performance of the same participants but at different ages. Furthermore, we recommend the use of more sensitive and elaborate (e.g., online) measures of statistical learning at both the individual and group levels. For example, our meta-analysis included only one ERP study (von Koss Torkildsen, 2010). Interestingly, the difference between people with and without SLI in this particular study was relatively high (see Figure 2.3). Potentially, the ERP measure compared with the accuracy measure is more sensitive in picking up differences in performance between people with and without SLI. We recommend future studies to further investigate this potential difference in a within-subject design with results of both measurement types for each individual.

2.5 Conclusion

In conclusion, the result of our meta-analysis shows that there is a moderate-to-large statistical learning deficit in people with SLI. This result is congruent with the hypothesis that people with SLI are less effective in statistical learning in the auditory verbal domain than people without language impairment. These results motivate the development of statistical learning–based interventions for children with SLI. More studies are needed, however, to perform more fine-grained analyses on the determinants of statistical learning deficiencies in the auditory verbal domain in people with SLI.

Acknowledgements

This work is part of the research programme Vernieuwingsimpuls – Vidi (examining the contribution of procedural learning to grammar and literacy in children) with project number 27689005, which is (partly) financed by the Dutch Research Council (NWO), and awarded to Judith Rispens (PI). We would like to thank Julia Evans, Eileen Haebig, Hsinjen Julie Hsu, Janne von Koss Torkildsen, and Agnes Lukács for providing additional data and information that enabled us to perform the meta-analysis. In addition, we would like to thank Sho Tsuji and Christina Bergmann for their feedback and advice and for sharing their thoughts on creating a Community-Augmented Meta-Analysis. Finally, we would like to thank Sascha Couvee for her help during the literature search.

Chapter 3

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

This chapter is a slightly modified version of the paper that was published as:

Lammertink, I., van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019). Auditory statistical learning in children: Novel insights from an online measure. *Applied Psycholinguistics*, 40(2). 279–302.

Data and scripts for analyses: https://osf.io/bt8ug/

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 present 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 times measures (perhaps in addition to the offline grammaticality judgments) to further investigate nonadjacent dependency learning in children.

3.1 Introduction

Statistical learning, the ability to detect structure in the environment, plays a key role in the development of language, perception, motor skills, and social behaviour (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, sensitivity 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, 2018). 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-aged 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 pseudoword 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 dependencies.

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 head-turn 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.

3.1.1 The present 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 dependencies, 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.

3.2 Method

3.2.1 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; months, range = 5;9–8;6 years; months). No hearing, vision, language, or behavioural 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.

3.2.2 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).

3.2.3 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 nonexistent 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 colour 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 twenty percent (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 3.1) 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 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 3.1). 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 3.1. 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.

Table 3.1 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

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 items 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 colour 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 nonadjacent dependency rule (incorrect; **tep wadim mip* in the example above).

Children were presented with sixteen utterance pairs. In eight 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 nonadjacent 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 see van Witteloostuijn, Lammertink, Boersma, Wijnen, & Rispens, 2019). 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, de Klerk and Wijnen (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, see van Witteloostijn, Lammertink et al., 2019), 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 3.1.

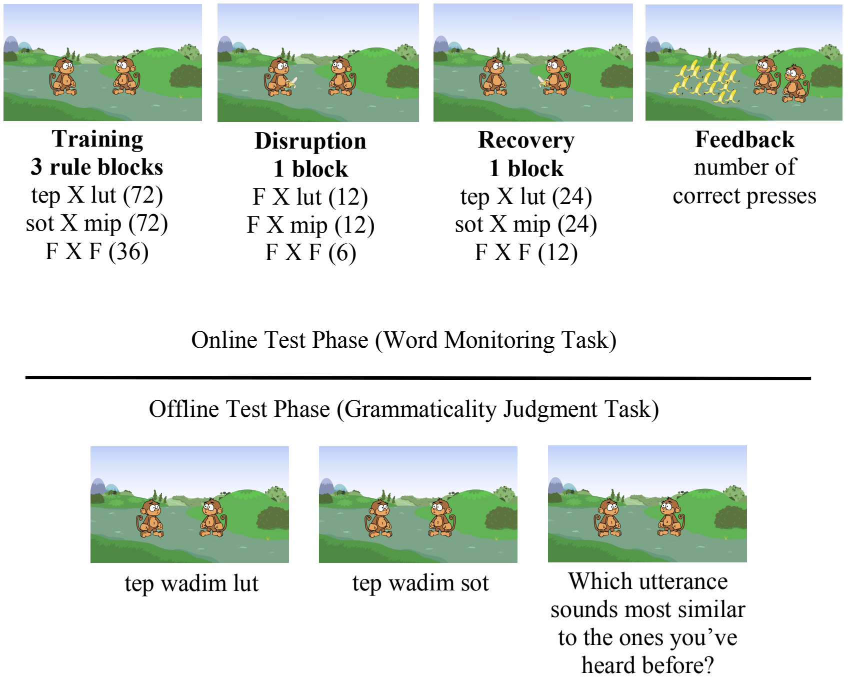


Figure 3.1 Visual representation of the online test phase (word monitoring task) and offline test phase (grammaticality judgment task) of the nonadjacent dependency learning task.

3.2.4 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[7]) 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 standard deviations (*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 (4464 observations) and used this data set to answer our research question.

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).

3.2.5 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

[7] Seventeen percent of target trials and 16% of nontarget trials, also the total percentage of errors, was approximately equally distributed across the five blocks.

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 or 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 peak) and their offline measure of learning (i.e. accuracy score). 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 r correlation coefficient. In addition, we made our data, data preprocessing script, 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 items 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 3.2.

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

3.3.1 Online measure (RTs)

Online measure: Descriptives. Mean RTs to the target items and nontarget items across the training blocks, disruption block, and recovery block are visualized in Figure 3.2. As we are interested in the learning trajectories across the third training block, disruption block, and recovery block, Table 3.3 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 "confirmatory disruption peak" model; Table 3.4). In order that our estimate of the effect of the first contrast ("DisruptionPeak") 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 (DisruptionPeak) estimated how much the true mean RT in the disruption block (which is coded as $+\frac{2}{3}$) exceeds the average of the true mean RT in the third training block (coded as $-\frac{1}{3}$) and the true mean RT in the recovery block (also coded as $-\frac{1}{3}$). This first contrast of the predictor Block (DisruptionPeak) intends to answer our specific research question (i.e., whether RTs are disrupted by removal of the nonadjacent dependency)[8]. 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; $+\frac{1}{2}$) were 52 ms faster than their negative responses (average RT nontarget items; $-\frac{1}{2}$; $t = -4.6$; p = .00003; 95% CI [−75, −30]), so we can conclude that children are generally faster in giving a positive than a negative response.

Table 3.3 Response times in milliseconds (ms) to the target items 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)

[8] The second contrast of Block ("PrePostDisruption") estimated how much the true mean RT in the recovery block ($+\frac{1}{2}$) exceeds the true mean RT in the third training block ($-\frac{1}{2}$). As this contrast does not directly answer our research question, we disregard the model outcome of this comparison.

Table 3.2 Summary of the confirmatory and exploratory research questions and operationalization of the predictors

Predictor	**Operationalization**	**Type of RQ**
Online response time measures		
DisruptionPeak	Do response times increase when NAD rules disappear?	Confirmatory
Targetness	Are response times to target items (positive responses) faster than response times to nontarget items (negative responses)?	Sanity check
DisruptionPeak × Targetness	Is the disruption peak different for target items than nontarget items?	Exploratory
DisruptionPeak × Age (days)	Does age modulate the size of the disruption peak?	Exploratory
DisruptionPeak × ExpVersion	Does experiment version modulate the size of the disruption peak?	Counterbalancing
Offline accuracy scores		
Intercept	Are accuracy scores different from chance?	Confirmatory
Generalization	Is there a difference in accuracy between familiar and novel items?	Exploratory
Age (days)	Does age modulate accuracy?	Exploratory
ExpVersion	Does experiment version modulate accuracy?	Counterbalancing

(*Table continues*)

Table 3.2 (*Continued*)

Predictor	**Operationalization**	**Type of RQ**
Disruption peak (online measure) and accuracy score (offline measure)		
Pearson *r* correlation	Are the size of the disruption peak and accuracy score correlated?	Exploratory
Awareness questionnaire		
	Have children explicit knowledge of the NAD rules?	Exploratory

Note. RQ = research question; NAD = nonadjacent dependency.

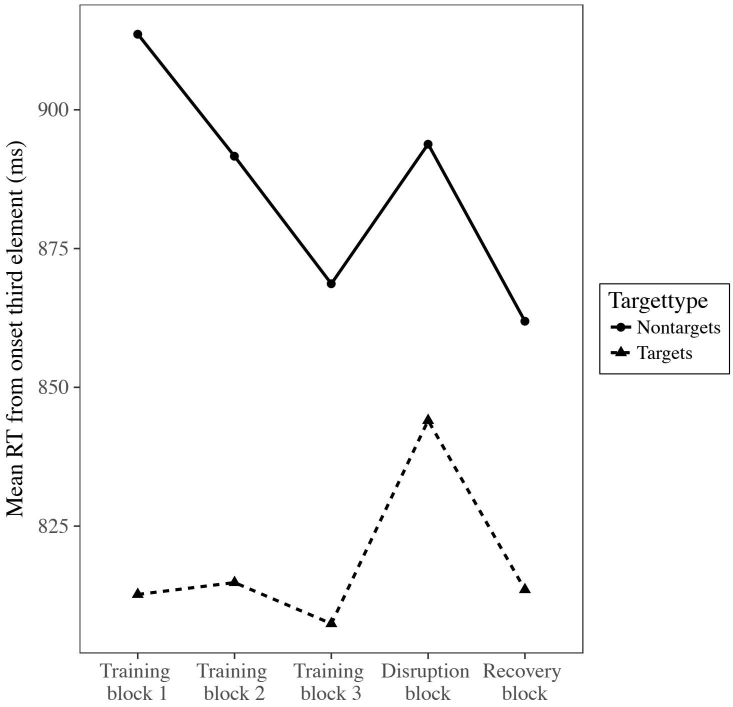


Figure 3.2 Participants' mean raw response times (RTs) across all five blocks of the online exposure phase for target items (triangle shaped and dashed line) and nontarget items (round shaped and solid line) separately. Please note that these raw RTs are only displayed for ease of exposition and that they do not represent the outcome of our confirmatory hypothesis testing. Therefore, (descriptive) differences in these raw RTs cannot be used to interpret the strength of the effects reported later in this paper.

Second, we explored whether the disruption peak differed between target items and nontarget items (interaction between DisruptionPeak 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 nontarget items. To further explore this null result, we fitted two additional models in which we re-referenced the contrast coding. To obtain a t value for the disruption peak in target items, the contrasts were set as target 0 (previously $-\frac{1}{2}$) and nontarget +1 (previously $+\frac{1}{2}$). 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.

Third, we checked whether the disruption peak differs between the two versions of the experiment (interaction between DisruptionPeak 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 re-referenced the model contrasts to obtain a *t* value for the disruption peak in experiment version 1 (version 1: 0; version 2: +1) and experiment version 2 (version 1: +1; version 2: 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 DisruptionPeak). 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.

3.3.2 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 3.3A. 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 3.3B.

Table 3.4 Outcome of the confirmatory disruption peak model (linear mixed-effects model on raw response times; 4464 observations)

***Random effects of subject* (*N* = 46)**	***SD* (ms)**
Intercept	80
DisruptionPeak	39
PrePostDisruption	25
Targetness	65
DisruptionPeak × Targetness	53
PrePostDisruption × Targetness	27
***Random effects of X-element* (*N* = 24)**	***SD* (ms)**
Intercept	92
ExpVersion	10

(Table continues)

Table 3.4 (*Continued*)

Fixed effect	***β* (ms)**	**95% Confidence interval (ms)**	***t***	***p***
Intercept	+864	[+819, +908]	+39	$7.4{\cdot}10^{-9}$
DisruptionPeak[a]	+36	[+17, +56]	+3.8	.00038
PrePostDisruption	+1	[−14, +15]	+0.069	.94
Targetness[b]	−52	[−75, −30]	−4.6	$3.0{\cdot}10^{-5}$
Age (days)	−13	[−37, +11]	−1.1	.29
ExpVersion	+94	[+45, +143]	+3.8	.0034
DisruptionPeak × Targetness[c]	+9	[−23, +41]	+0.55	.58
PrePostDisruption × Targetness	+13	[−14, +40]	+0.97	.33
DisruptionPeak × Age (days)[c]	−5	[−24, +14]	−0.51	.61
PrePostDisruption × Age (days)	−5	[−21, +10]	−0.70	.48
Targetness × Age (days)	+7	[−16, +30]	+0.60	.55

(*Table continues*)

Table 3.4 (*Continued*)

Fixed effect	**B (ms)**	**95 % Confidence interval (ms)**	***t***	***p***
DisruptionPeak × ExpVersion[d]	+6	[−31, +43]	+0.33	.75
PrePostDisruptionPeak × ExpVersion	−4	[−34, +26]	−0.25	.80
Targetness × ExpVersion	+68	[+22, +113]	+3.0	.0043
Age (days) × ExpVersion	−34	[−83, +15]	−1.4	.17
DisruptionPeak × Targetness × Age (days)	+21	[−13, +55]	+1.2	.22
PrePostDisruption × Targetness × Age (days)	+8	[−20, +35]	+0.54	.59
DisruptionPeak × Targetness × ExpVersion	+37	[−28, +102]	+1.2	.25
PrePostDisruption × Targetness × ExpVersion	−6	[−60, +47]	−0.23	.82.
DisruptionPeak × Age (days) × ExpVersion	+20	[−18, +58]	+1.1	.29

(*Table continues*)

Table 3.4 (*Continued*)

Fixed effect	**β (ms)**	**95 % Confidence interval (ms)**	***t***	***p***
PrePostDisruption × Age (days) × ExpVersion	+8	[−23,+39]	+0.52	.60
Targetness × Age (days) × ExpVersion	+41	[−5,+87]	+1.8	.080
DisruptionPeak × Targetness × Age (days) × ExpVersion	−12	[−81,+55]	−0.36	.72
PrePostDisruption × Targetness × Age (days) × ExpVersion	+10	[−46,+65]	+0.35	.73

Note. [a]Relevance is confirmatory; [b]Relevance is data check; [c]Relevance is exploratory; [d]Relevance is counterbalancing.

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 3.3A; Table 3.5). The predictor Generalization estimated by what ratio the children scored better on items with a familiar X-element ($+\frac{1}{2}$) than on items with a novel X-element ($-\frac{1}{2}$). 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 familiar items ($z = -0.66$; $p = .51$, 95% CI [0.65, 1.25]), 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: $-\frac{1}{2}$; version 2: $+\frac{1}{2}$) and Age. The model estimates of both predictors were not significantly different from 1 (Table 3.5), and therefore the results do not generalize to the population.

3.3.3 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 3.4). 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 DisruptionPeak 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.

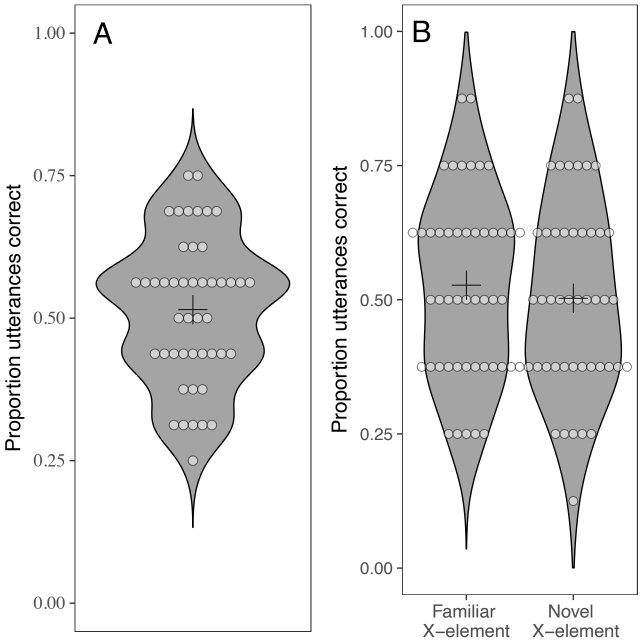


Figure 3.3 Descriptive visualizations of distribution of (A) the overall mean correctness probabilities on the two-alternative grammaticality judgment task and (B) the mean correctness probabilities by generalization. The dots represent the individual scores, and the cross indicates the overall group mean. Please note that we did not obtain these correctness probabilities from the statistical model. These descriptive data are only displayed for ease of exposition and do not represent the outcome of the generalized linear mixed-effects model. Therefore, (descriptive) differences in this plot cannot be used to interpret the strength of the effects reported later in this paper.

Relationship online measure and offline measure: Exploratory results. The Pearson *r* correlation coefficient[9] between children's online measure of disruption and their offline measure of learning was not statistically significantly different from zero ($r = -.17$; $p = .27$; Figure 3.4). Therefore, we have no evidence that children's online disruption score correlates with their offline accuracy score.

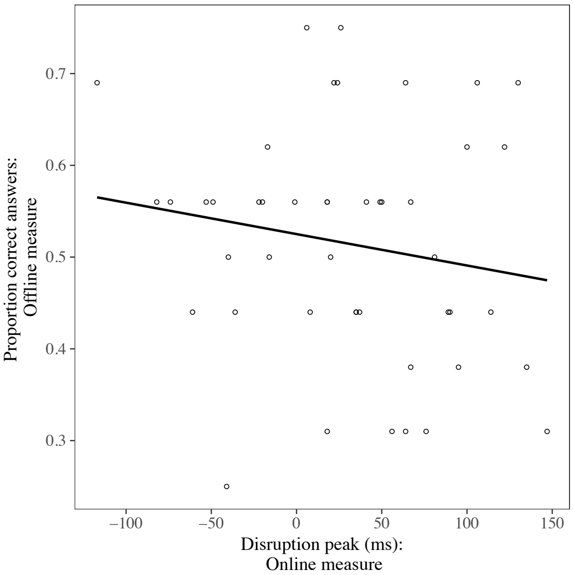


Figure 3.4 Scatter plot and regression line that represents the descriptive association between children's individual online disruption score (x-axis) and children's individual accuracy score on the grammaticality judgment task (y-axis).

[9] Note that this correlation does not consider the between-subject variable ExpVersion. In an alternative analysis, we added children's offline learning scores to the linear mixed-effects confirmatory disruption peak model (OfflinePlus model) and compared this OfflinePlus model to the confirmatory disruption peak (Table 3.4) by means of the *analysis of variance* function in R. When comparing both models, the OfflinePlus model did not significantly improve the confirmatory disruption peak model ($\chi2 = 1.74$; $p = .19$). 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.

Table 3.5 Outcome for the offline measure of nonadjacent dependency learning (generalized linear mixed-effects model on correctness probability; 736 observations)

Random effects of subject ($N = 46$)				***SD* (log odds)**	
Intercept				0.058	
Generalization				0.058	
Random effects of X-element ($N = 16$)					
Intercept				0.11	
ExpVersion				0.35	
Fixed effect	**β_{model} (log odds)**	**$\beta_{\text{transformed}}$ (odds)**	**95% Confidence interval (odds)**	***z***	***p***
Intercept[a]	+0.064	1.07	[0.90,1.26]	+0.81	.42
Generalization[a]	−0.10	0.90	[0.65,1.25]	−0.66	.51
Age (days)[b]	−0.011	0.99	[0.84,1.15]	−0.14	.89
ExpVersion[c]	−0.047	0.95	[0.66,1.36]	−0.27	.78

(*Table continues*)

Table 3.5 (*Continued*)

Fixed effect	$\boldsymbol{\beta}_{\text{model}}$ **(log odds)**	$\boldsymbol{\beta}_{\text{transformed}}$ **(odds)**	**95% Confidence interval (odds)**	z	$\boldsymbol{p}$
Generalization × Age (days)	+0.078	1.08	[0.79,1.46]	+0.51	.61
Generalization × ExpVersion[c]	−0.039	0.96	[0.47,1.97]	−0.11	.91
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

Note. [a]Relevance is confirmatory; [b]Relevance is exploratory; [c]Relevance is counterbalancing.

3.3.4 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.

3.4 Discussion

The present 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 the 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 measures 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; Isbilen 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 metalinguistic or explicit decision (Franco et al., 2015). This metalinguistic 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 present study have been argued to be psychometrically weak for measuring individual statistical learning performance (Siegelman, Bogaerts, & Frost 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, Bogaerts, & Frost, 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 items 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 [Chapter 2 of this dissertation]; Obeid, Brooks, Powers, Gillespie-Lynch, & Lum, 2016). In these studies, we see that in people with DLD 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, 2019).

3.5 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 work is part of the research programme Vernieuwingsimpuls – Vidi (examining the contribution of procedural learning to grammar and literacy in children) with project number 27689005, which is (partly) financed by the Dutch Research Council (NWO), and awarded to Judith Rispens (PI). 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).

Chapter 4

Children with developmental language disorder have an auditory verbal statistical learning deficit: Evidence from an online measure

This chapter is a slightly modified version of the paper that was published as:

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Children with developmental language disorder have an auditory verbal statistical learning deficit: evidence from an online measure. *Language Learning*, *70*(1), 137–178.

Publicly accessible 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

Data, materials and scripts for analyses: https://osf.io/8a3yv/

Abstract

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; Mean age = 9.1 years) and without DLD (*N* = 36; Mean 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 their 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.

4.1 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 (verbal) input (Evans, Saffran, & Robe-Torres, 2009; Hsu & Bishop, 2014a; Lammertink, Boersma, Wijnen, & Rispens, 2017 [Chapter 2 of this dissertation]; 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), which is 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.

4.1.1 Background literature

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).

4.1.2 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 [Chapter 2 of this dissertation]; 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 11), 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 crossed 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 modelled 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, researchers have not 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, 2019 [Chapter 3 of this dissertation]). 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 (2018) 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, under review [Chapter 6 of this dissertation]). 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, 2016a; 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.

4.1.3 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, Rodrìgez-Fornells, & de Diego-Balaguer, 2016; Misyak et al., 2010; Vuong, Meyer, & Christiansen, 2015) and in primary-school-aged children (Lammertink, van Witteloostuijn et al., 2019 [Chapter 3 of this dissertation]). 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' behaviour 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).

4.1.4 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 behaviour in an offline forced-choice task.

4.2 Method

4.2.1 Participants

We recruited 37 children with DLD and 59 typically developing children aged between 7 and 11 years to participate in our study[10]. At the end of the study, we

[10] 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, Rispens and Wijnen (2020 [Chapter 5 of this dissertation]) and Lammertink, Boersma, Wijnen and Rispens (under review [Chapter 6 of this dissertation]). Van Witteloostuijn, Boersma, Wijnen, and Rispens (2019a, 2019b, submitted) have also described

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.

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. Typically developing children were enrolled on an opt-out basis. Table 4.1 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).

4.2.2 Recruitment and inclusion of children with DLD

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 (FOSS/ Stichting Hoormij), and through self-employed speech therapists. All participants in this group had been diagnosed with DLD by licensed clinicians

a subset of the typically developing participants in separate studies, with different research questions, and a different clinical group (developmental dyslexia).

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 Progressive Matrices subtest (Raven, Raven, & Court, 2003), a standardized measure of nonverbal 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 nonverbal intelligence criterion (and thus a Raven Progressive Matrices score of at least 17%). Only shortly thereafter, Bishop et al. (2017) made their recommendation that low nonverbal intelligence 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.

4.2.3 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 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 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

Table 4.1 Summary of the group characteristics

	DLD (*N* = 36)	**TD (*N* =36)**	**Difference DLD – TD**		
	Mean [Range]	**Mean [Range]**	***t***	***p***	**95% CI**
Age (years; moths)	9;1 [7;8, 10;4]	9;1 [7;8, 10;4]	+0.032	.97	[−0; 3, +0; 3]
Nonverbal intelligence					
Raw	36 [23, 49]	36 [26, 55]	+0.019	.98	[−3, +3]
Standardized (percentiles)	63 [17, 96]	64 [20, 98]			
Socioeconomic status	+0.22 [−2.57, +2.09]	−0.06 [−1.28, +1.15]	+1.2	.23	[−0.18, +0.75]

Note. TD = typically developing.

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.

4.2.4 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., 2019 [Chapter 3 of this dissertation], 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 4.1). In all utterances, Element 1 was a monosyllabic Dutch pseudoword (e.g., *tep*), Element 2 was a bisyllabic Dutch pseudoword (e.g., *wadim*), and Element 3 was again a monosyllabic Dutch pseudoword (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 4.2 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 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.

Table 4.2 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

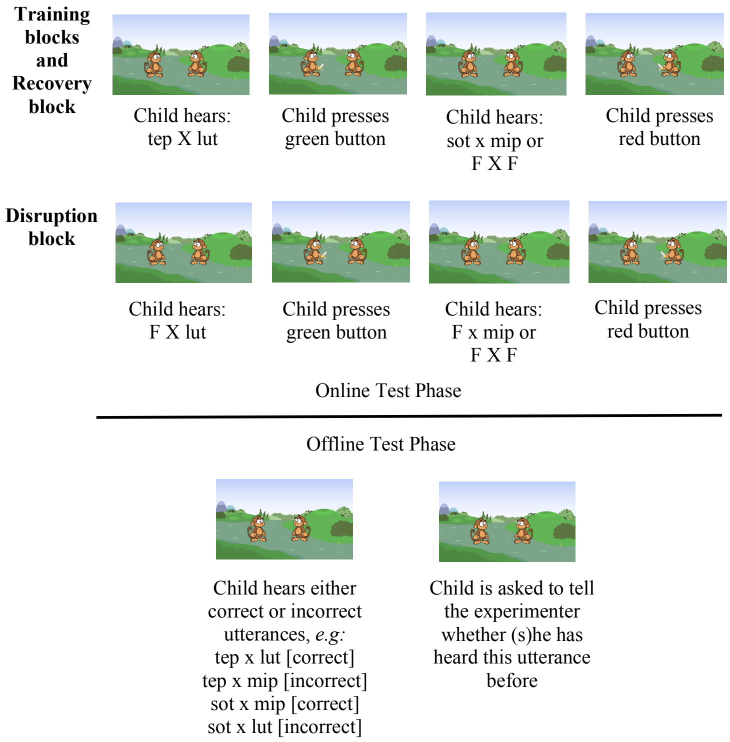

Training blocks and Recovery block
Child hears: tep X lut
Child presses green button
Child hears: sot x mip or F X F
Child presses red button
Disruption block
Child hears: F X lut
Child presses green button
Child hears: F x mip or F X F
Child presses red button
Online Test Phase
Offline Test Phase
Child hears either correct or incorrect utterances, e.g.: tep x lut [correct] tep x mip [incorrect] sot x mip [correct] sot x lut [incorrect]
Child is asked to tell the experimenter whether (s)he has heard this utterance before

Figure 4.1 Example of the nonadjacent dependency task. In this example, the green button is on left and the red button is on right. If a child presses the green button, the banana thus appears on left and if a child presses the red button, the banana appears on right.

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 4.2). 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 4.3 provides a short description of each measure.

4.2.5 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 (see Chapter 5 of this dissertation), or a serial reaction time task (see chapter 6 of this dissertation) – and was then followed 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 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, Rispens, and Wijnen (2020 [Chapter 5 of this dissertation]) and Lammertink, Boersma, Wijnen, and Rispens (under review [Chapter 6 of this dissertation]). 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 ($N = 14$).

Table 4.3 Description of other cognitive and language measures used in the study

Task	**Description**	**Possible range (raw scores)**
Raven's Progressive Matrices (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

4.2.6 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 analysed 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 (non-target, target), and ExpVersion (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 ExpVersion, 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 A4), 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 ExpVersion[11]. 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., 2019 [Chapter 3 of this dissertation]). 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

[11] 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 ExpVersion 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.

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 as $-\frac{1}{2}$) would allow us to answer our confirmatory research question. The predictor variables ExpVersion, 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 4.3). 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 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 4.2), one anonymous reviewer asked us to explore whether the difference between participants' response times in the first training block and their response times in the third training block (i.e., the response time gain) was larger for typically developing participants than for participants with DLD. In exploring this first issue, we analysed 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 see 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 the first training block to the third training block, 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 DisruptionPeak (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 DisruptionPeak 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 behaviour 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 ExpVersion (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 ExpVersion (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 A4). We assessed the statistical significance of the predictors using 95% Wald confidence intervals.

4.3 Results

4.3.1 Background measures: group comparisons on the cognitive and language tasks

Table 4.4 presents the raw scores and, when available, the standardized norm or percentile scores for the cognitive and language tasks (described in Table 4.3) for both groups. Between-group *t* tests (see Table 4.4) 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.

4.3.2 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., 2019 [Chapter 3 of this dissertation]). Responses were coded as incorrect if participants pressed the wrong button colour 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 4.2). 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 on.

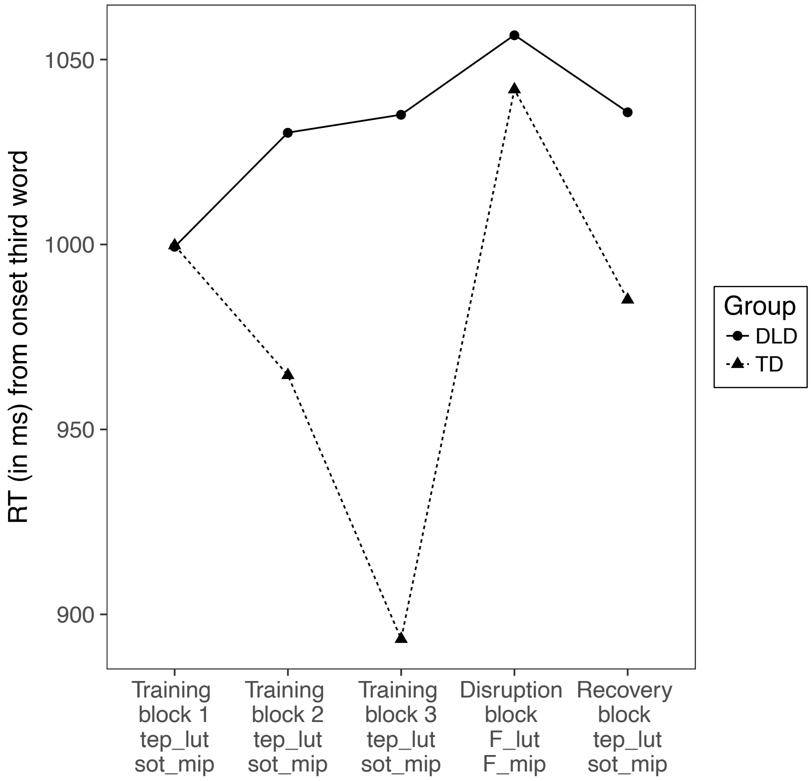


Figure 4.2 Participants' mean raw response times (RTs) across all five blocks of the online exposure phase. DLD (round shape and solid line); TD = typically developing (triangle shaped and dashed line). Please note that these raw RTs are only displayed for ease of exposition and that they do not represent the outcome of our confirmatory hypothesis testing. Therefore, (descriptive) differences in these raw RTs cannot be used to interpret the strength of the effects reported later in this paper.

Table 4.4 Children,s raw and – when available – norm scores or percentiles for the background measures and group comparisons

	DLD ($N = 36$)	**TD ($N = 36$)**	**Difference DLD – TD**		
	Mean [Range]	**Mean [Range]**	***t***	***p***	**95% CI**
Verbal short-term memory					
Raw	6 [3, 9]	9 [6, 12]	−7.7	$6.1 \cdot 10^{-11}$	[−3, −2]
Standardized (norm scores)	6[a] [1[a], 12]	11 [6[a], 15]			
Verbal working memory					
Raw	3 [2, 5]	4 [2, 8]	−3.4	.0011	[−2, −4]
Standardized (norm scores)	8 [4[a], 12]	10 [5[a], 16]			
Sustained attention					
Raw	7 [1, 10]	8 [3, 10]	−0.8	.44	[−1, +1]
Standardized (norm scores)	8 [1[a], 13]	9 [3[a], 13]			

(*Table continues*)

Table 4.4 (*Continued*)

	DLD ($N = 36$)	TD ($N = 36$)	Difference DLD – TD		
	Mean [Range]	**Mean [Range]**	***t***	***p***	**95% CI**
Receptive vocabulary knowledge					
Raw	101 [78, 118]	115 [98, 140]	−5.8	$1.6 \cdot 10^{-7}$	[−18, −9]
Standardized (percentiles)	33 [1^{a}, 84]	63 [6^{a}, 95]			
Expressive morphosyntactic proficiency					
Raw	31 [12, 67]	59 [32, 81]	−9.2	$1.1 \cdot 10^{-13}$	[−35, −22]
Standardized (norm scores)	5^{a} [1^{a}, 13]	11 [3^{a}, 16]			
Expressive morphological proficiency					
Raw	22 [12,29]	28 [22,30]	−7.4	$2.2 \cdot 10^{-9}$	[−8, −4]

Note. TD = typically developing; F = Female; M = Male. CI = confidence interval. aStandardized 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).

4.3.3 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, Δ*z* = 0.19, *t* = 2.23, 95% profile CI [0.02, 0.36], *p* = .03 (see also Table 4.5 and Figure 4.3), 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 × DisruptionPeak interaction, we fitted two additional 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, Δ*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, Δz = 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 *r* correlation between the size of participants' individual disruption peak for even items (random slopes for the predictor DisruptionPeak 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].

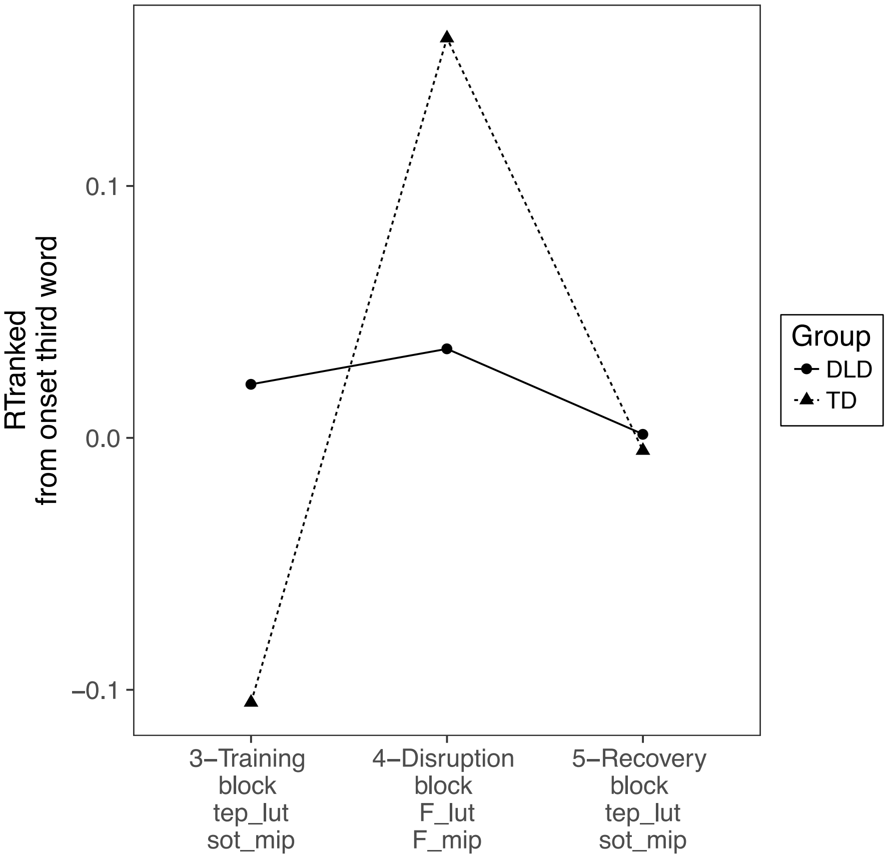


Figure 4.3 Interaction between the size of the disruption peak and the predictor variable group. RT = response time; TD = typically developing.

4.3.4 Online measure: Exploratory results

From the visualization of participants’ raw response times across the five blocks (Figure 4.2), 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 DisruptionPeak (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.

Table 4.5 Outcome of the relevant estimates from the confirmatory disruption peak model (linear mixed-effects model on normalized response times; 8015 observations)

Random effects of subject ($N = 72$)		**Correlation**		
	***SD* (Δ *z*)**	**Intercept**	**DisruptionPeak**	**PrePostDisruption**
Intercept	0.35			
DisruptionPeak	0.16	−.31		
PrePostDisruption	0.24	−.20	+.45	
Targetness	0.11	+.61	−.32	−.07
Random effects of X-element ($N = 24$)				
	***SD* (Δ *z*)**	**Intercept**	**ExpVersion**	
Intercept	0.31			
ExpVersion	0.04	−.75		
Group	0.07	+.36	+.35	
Residual *SD*	0.84			
Fixed effect	**β (Δ *z*)**	**95% CI (Δ *z*)**	***t***	***p***
DisruptionPeak × Group[a]	+0.19	[+0.02, +0.36]	+2.23	.03

Note. CI = confidence interval. The full model outcome (including all fixes effects) can be found in the R markdown script at https://osf.io/8a3yv. An operationalization of the predictors can be found in appendix A4

4.3.5 Further exploration of the link between online statistical learning and grammatical proficiency

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

4.3.6 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 4.6 and Figure 4.5). Therefore, we had no evidence that our offline measure captured children's sensitivity to the NADs.

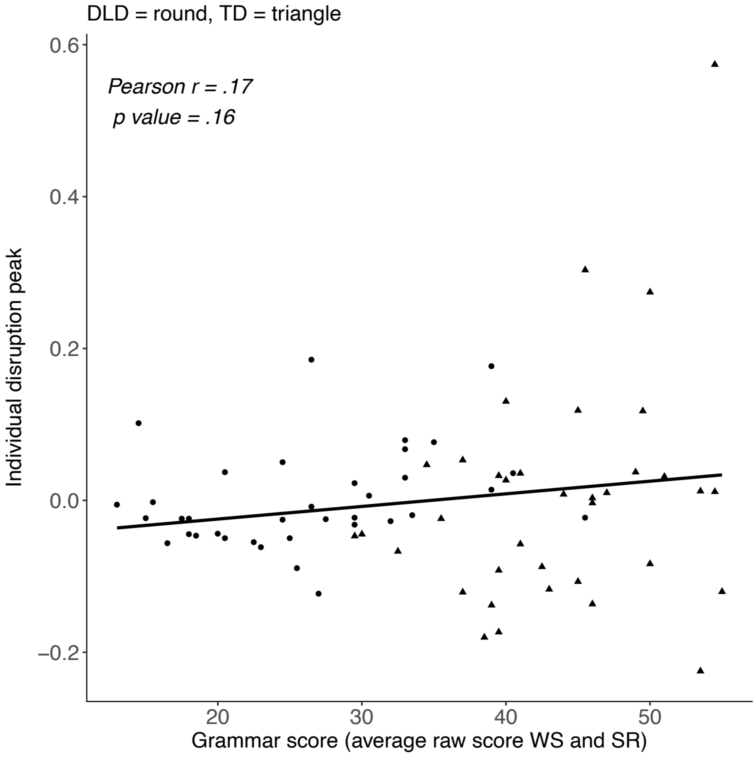


Figure 4.4 Graphic (descriptive) representation of the relationship between participants' individual disruption peaks and their grammar performance. TD = typically developing; WS = word structure; SR = sentence recall. Individual disruption peaks (random slope in Δz, controlled for Group).

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 4.6). Therefore, we could not conclude that the Rule effect differed between children with DLD and typically developing children.

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 4.6). 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 4.6).

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 4.6). 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 4.6). Our task instructions might have caused this familiarity effect, however (see Discussion).

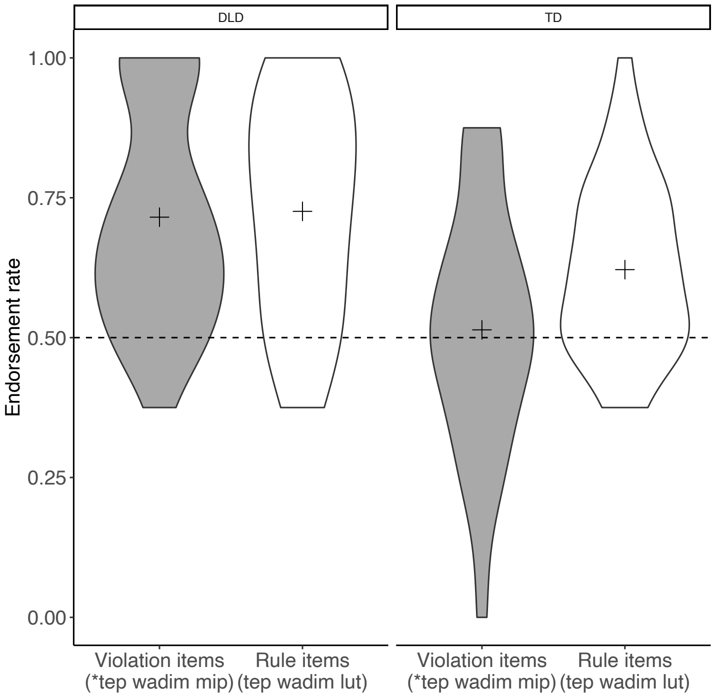


Figure 4.5 Graphic (descriptive) representation of endorsement rates for item types by group. DLD = developmental language disorder; TD = typically developing. Please note that we did not obtain these endorsements rates from the statistical model. These descriptive data are only displayed for ease of exposition and do not represent the outcome of the generalized linear mixed model.

Table 4.6 Outcome of the relevant estimates for the offline measure of nonadjacent dependency learning (generalized linear mixed-effects model on endorsement rate; 1152 observations)

Random effects of subject (*N* =72)		**Correlation**		
	***SD* (log odds)**	**Intercept**	**Rule**	**Generalization**
Intercept	0.58			
Rule	0.54	−.42		
Generalization	1.02	−.27	+.60	
Rule × Generalization	0.69	+.86	−.26	+.23
Random effects of X-element (*N* = 16)				
	***SD* (log odds)**	**Intercept**	**Group**	**ExpVersion**
Intercept	0.44			
Group	0.23	+.17		
ExpVersion	0.91	−.20	+.75	
Group × ExpVersion	0.56	−.18	−.68	−.91

(*Table continues*)

Table 4.6 (*Continued*)

Fixed effect	$\boldsymbol{\beta}_{\text{model}}$ **(log odds)**	$\boldsymbol{\beta}_{\text{transformed}}$ **(odds)**	**95% Confidence interval (odds)**	z	p
Intercept (yes bias)[a]	+0.79	2.2	[1.5,3.1]	+4.48	$7.5 \cdot 10^{-6}$
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, and odds ratios ratios. The full model outcomes (including all predictors) can be found in the R markdown script at https://osf.io/8a3yv. An operationalization of the predictors in Appendix A4. [a]Relevance is exploratory.

4.4 Discussion

4.4.1 A small auditory verbal 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 [Chapter 2 of this dissertation]), 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 [Chapter 2 of this dissertation]), 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 [Chapter 2 of this dissertation], 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.

4.4.2 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 (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. (2019 [Chapter 3 of this dissertation]) 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.

4.4.3 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 4.2) 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.

4.5 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 [Chapter 2 of this dissertation]), 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.

Acknowledgements

This work is part of the research programme Vernieuwingsimpuls – Vidi (examining the contribution of procedural learning to grammar and literacy in children) with project number 27689005, which is (partly) financed by the Dutch Research Council (NWO), and awarded to Judith Rispens (PI). We extend our gratitude to all children who participated and to their parents, teachers, and speech therapists who facilitated the children's participation. More specifically, we would like to thank Viertaal special education; The Royal Dutch Auris Group; The Royal Dutch Kentalis; Pento, and FOSS/ stichting Hoormij (Dutch parents' association for parents of children with DLD) for their help in the recruitment of children with DLD. Also, we would like to thank the four primary schools that participated. Finally, we thank Iris Broedelet-Resink, Sascha Couvee, Darlene Keydeniers, Maartje Hoekstra (test assistants), Dirk Jan Vet (technical implementation of the experiment), and Merel van Witteloostuijn (extensive help with the design of the experiment).

Chapter 5

Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD

This chapter is a slightly modified version of the paper that was accepted for publication as:

Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing: An Interdisciplinary Journal*. Advance online publication.

Data and scripts for analyses: https://osf.io/8gpjt/

Abstract

Visual statistical learning (VSL) has been proposed to underlie literacy development in typically developing children. A deficit in VSL may thus contribute to the observed problems with written language in children with dyslexia. Interestingly, although many children with developmental language disorder (DLD) exhibit problems with written language similar to those seen in children with dyslexia, few studies investigated the presence of a VSL deficit in DLD, and we know very little about the relation between VSL and literacy in this group of children. After testing 36 primary-school-aged children (ages 7;8 – 10;4) with DLD and their typically developing peers on a self-paced VSL task, two reading tasks and a spelling task, we find no evidence for or against a VSL deficit in DLD, nor for associations between VSL and literacy in DLD. We discuss the implications for our understanding of language (and literacy) difficulties in children with DLD.

5.1 Introduction

Language therapists, clinical linguists and scientists who work with children with developmental language disorder (DLD) have long been interested in understanding the cognitive mechanisms underlying the language problems seen in these children. By definition, children with DLD have deficits in language that cannot be attributed to neurological damage, hearing impairment, intellectual disability, or unfavourable psychosocial/educational conditions. The difficulties with language manifest themselves across multiple areas such as the lexicon, morphology, (morpho)syntax, discourse (Leonard, 2014), reading (McArthur, Hogben, Edwards, Heath, & Mengler, 2000) and spelling (Joye, Broc, Olive, & Dockrell, 2019). Also, they frequently co-occur with difficulties in other cognitive domains such as attention, working memory (e.g., Ebert & Kohnert, 2011, Montgomery, Evans, & Gilliam, 2018) and motor skills (Hill, 2001). This wide range of observed difficulties makes it difficult to point to a core underlying (cognitive) deficit for the disorder and thus far the observed language problems in children with DLD have been explained from language-specific deficits (see Leonard, 2014, chapter 9 for an overview) as well as from deficits in more general learning or processing mechanisms that contribute to language development (e.g., auditory perception deficits: Tallal, Stark, & Mellits, 1985; slower processing (of spoken language): Miller, Kail, Leonard, & Tomblin, 2001; limited short-term memory and working memory capacities: Archibald, & Gathercole, 2006; Montgomery et al., 2018). In the present paper, we seek evidence for one of these more general accounts, namely that the problems observed in children with DLD stem from a general cognitive statistical learning deficit (Evans, Saffran & Robe-Torres, 2009; Hsu and Bishop, 2014a; Lammertink, Boersma, Wijnen & Rispens, 2017 [Chapter 2 of this dissertation], Obeid, Brooks, Powers, Gillespie-Lynch & Lum 2016; Wijnen, 2013). Before we turn into explaining why the study of visual statistical learning in DLD is interesting, we first outline how sensitivity to structural regularities in the input (i.e., statistical learning) may play a role in children's language development.

5.1.1 Language learning through statistics

Natural languages reflect structural regularities at the sound, word and sentence level. The ability to detect and learn these regularities may be crucial for language development as it has been proposed to underlie word segmentation (Saffran &

Graf Estes, 2006) and the construction of linguistic categories and dependencies (e.g., Mintz, 2003; Wijnen, 2013). Indeed, there seems to be a predictive relation between detecting and learning regularities from verbal input (statistical learning) and different aspects of language (e.g., vocabulary knowledge: e.g., Spencer, Kaschak, Jones, & Lonigan, 2015; Shafto, Conway, Field, & Houston, 2013; morphology/grammar: Hamrick, Lum, & Ulman, 2018 and syntactic processing: Kidd, 2012; Kidd & Arciuli, 2016; Wilson et al., 2018). Another source of evidence for a link between statistical learning and language ability comes from studies in people with DLD: these studies have shown that people with DLD are less sensitive to statistical regularities in auditorily presented verbal stimuli than people without DLD (meta-analyses: Lammertink et al., 2017 [Chapter 2 of this dissertation], Obeid et al., 2016). In these studies participants typically listen to a continuous stream of auditorily presented nonsense syllables, either presented in a continuous manner (e.g., *bupadadutaba*; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997) or with short pauses in between (e.g. *tep wadim lut;* Gómez, 2002). Unbeknowst to the participants the nonsense syllables form words (the example above consists of two words: *bupada* and *dutaba*) or their order of appearance in the utterance is governed by rules (in the second example above, *tep* and *lut* always co-occur). These words and rules can be learned if participants are sensitive to the transitional probabilities or nonadjacent dependencies that underlie them. When people with and without DLD are tested on their knowledge of these words and rules, it has been shown that people without DLD outperform people with DLD. Hence, people with DLD show an auditory verbal statistical learning deficit as compared to people without DLD (see also Chapter 4 of this dissertation).

5.1.2 Statistical learning outside the language domain

Structure is not unique to language, however (e.g., "like language, music can be viewed as a system of structure regularities"; Leonard, 2014, p. 213), and therefore it has been hypothesised that humans may have a domain-general statistical learning mechanism. The hypothesis that a domain-general statistical learning mechanism, rather than a domain-specific learning mechanism (i.e., sensitivity to statistical patterns solely in the verbal input), is important for successful language acquisition, leads to two predictions. First, one would expect to observe correlations between people's ability to detect statistical regularities in other domains than language and their performance on language tasks. Second, the hypothesis also predicts that the statistical learning deficit observed in children with DLD is domain-general and should thus also be present outside the auditory verbal domain. As for the first prediction, there is evidence that in typically developing children and in children with dyslexia, statistical learning of regularities between nonverbal elements in the visual domain (e.g., unfamiliar cartoonlike characters, meaningless shapes or symbols) and visuomotor domain (e.g., a sequence of computer screen locations in which a cartoon or shape appears) correlates with reading performance (Arciuli & Simpson, 2012; Hedenius et al., 2013; Steacy et al., 2019; Vakil, Lowe, & Goldfus, 2015; van der Kleij, Groen, Segers, & Verhoeven, 2018; von Koss Torkildsen, Arciuli, & Wie, 2019) and grammatical proficiency (meta-analysis by Hamrick et al., 2018). As for the second prediction, there is also evidence that children with DLD perform worse on statistical learning tasks with nonverbal stimuli in the visuomotor domain that typically developing children (Lum, Conti-Ramsden, Morgan, & Ullman, 2014). Such a visuomotoric nonverbal statistical learning deficit has also been observed in children with dyslexia (Lum, Ullman, & Conti-Ramsden, 2013), but see recent studies reporting no evidence for or against such a deficit in dyslexia: Henderson, & Warmington (2017); van der Kleij et al. (2018); Schmalz, Altoè, & Mulatti (2017). Children with dyslexia also perform more poorly in their detection of nonverbal regularities (geometrical shapes or unfamiliar symbols) in the visual domain, hence they show a visual statistical learning (VSL) deficit (Pavlidou & Williams, 2014; Sigurdardottir et al., 2017). In this light, it should also be noted, however, that two different research groups concluded that the magnitude of the VSL deficit in dyslexia may be inflated as a result of publication bias (Schmalz et al., 2017; van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017).

5.1.3 A visual nonverbal statistical learning deficit in DLD

Interestingly, while there are some studies on VSL in children with dyslexia, studies on VSL in children with DLD are scarce. To the best of our knowledge only one study has thus far used a nonverbal VSL task to compare children with and without DLD (Noonan, 2018). Noonan found no evidence for or against a difference in VSL performance between children with and without DLD, but note that neither of the groups in her study showed evidence of learning or not learning the nonverbal regularities. Thus, it is still unknown whether the difficulties with language (and literacy) in children with DLD relate to a VSL deficit.

Investigating visual nonverbal statistical learning abilities in children with DLD is important for several reasons. Firstly, in the statistical learning literature on typical learners it has recently been claimed that – as opposed to being fully domain-general – the statistical learning mechanism is in part domain- or stimulus dependent (Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018). More specifically, Siegelman et al. observed a dissociation between people's statistical learning of verbal materials versus their learning of nonverbal materials. From this, Siegelman et al. claim that differences in prior knowledge of statistical structure may impact on performance in verbal statistical learning tasks differently from performance in nonverbal statistical learning tasks. They argue that the "tabula rasa*"* assumption (i.e., that learners have no expectations or prior knowledge regarding the underlying statistical structure) holds for nonverbal tasks but not for verbal tasks. With verbal materials, participants may always have expectations of the underlying structure based on their native language experience (Siegelman et al. refer to this as "linguistic entrenchment*"*). If this claim is true, this may mean that children with DLD are worse in detecting statistical regularities in verbal materials than their typically developing peers, not because they are less sensitive to the statistical regularities, but because they have less expectations of the underlying structure due to their language deficit. Only if the children with DLD also show a deficit in their detection of regularities in a nonverbal statistical learning task, one could conclude that reduced sensitivity to domain-general structural regularities contributes to the observed language problems in this group of children.

Secondly, on the basis of studies with typically developing children and those with children with dyslexia, it has been claimed that visual and visuomotoric statistical learning of nonverbal materials relates to literacy skills. While children learn to read and write, they need to detect which graphemes correspond to which phonemes and vice versa. In many orthographies, graphemes may correspond to multiple phonemes. Which phoneme should be used is then dependent on the context in which it appears. For example, in English, the grapheme 'c' may correspond to either /k/ as in *can't* or to /s/ as in *cent.* The statistical regularity to be learned is that the vowel that follows the 'c' determines its phoneme. When 'c' is followed by 'a', 'o' or 'u' it is usually pronounced as /k/; when it is followed by 'e', 'i' or 'y' it is usually pronounced as /s/. Children may use a statistical learning mechanism to detect these (context dependent) regularities in grapheme–phoneme correspondences (Arciuli, 2017, 2018; Treiman, 2018). Interestingly, children with DLD exhibit large individual differences in literacy performance; approximately half of the children with DLD have problems with reading and/or spelling (McArthur et al., 2000). In the present study, we will explore whether these large individual differences in literacy performance among children with DLD can be explained by individual differences in visual statistical learning – as has also been claimed for typically developing children and children with dyslexia.

Thirdly, a methodological reason for conducting the present study is that the evidence for a domain-general nonverbal statistical learning deficit comes mostly from studies using the serial reaction time task (Nissen & Bullemer, 1987). Although the serial reaction time task is widely used as a measure of people's visuomotoric nonverbal statistical learning ability, the validity of the task has been questioned (West, Vadillo, Shanks, & Hulme, 2017). Also, children with DLD often have subtle motor deficits (Hill, 2001) that may impact on their performance on this visuomotoric task. In the present study, we therefore use a nonverbal statistical learning task in the visual (rather than visuomotoric) domain to investigate the domain-generality of the statistical learning deficit in DLD. The reliability of the VSL task has been questioned as well, but recent modifications to the setup of the task are promising and seem to detect learning – in both adults (Siegelman, Bogaerts, & Frost, 2017; Siegelman, Bogaerts, Kronenfeld, & Frost, 2018) and children (van Witteloostijn, Lammertink, Boersma, Wijnen, & Rispens, 2019).

Fourthly, the present study follows one of the research directions put forward in Arciuli and Conway (2018). In this review paper, Arciuli and Conway conclude that it is important to further investigate under what conditions children with developmental disabilities can and cannot learn statistical regularities. As outcomes of studies like the present study may identify relative strengths and weaknesses of these children, they may be helpful in developing intervention studies that aim to support language learning in children with language difficulties.

5.1.4 The visual statistical learning paradigm

In the present study, we use a triplet learning paradigm to investigate children's sensitivity to statistical regularities in the visual nonverbal domain. In this paradigm, participants are visually exposed to a sequence of individual nonverbal elements (unique cartoon drawings or meaningless shapes) that appear one by one on a computer screen. Unbeknownst to the participants, the individual elements are distributed into fixed groups of three (triplets). Within these triplets, the transitional probability (TP) between elements is 1.0, but across triplets the TP is lower. After exposure to a series of elements, participants' knowledge of the triplets is assessed with an offline recognition test. Several research groups raised concerns on the use of a recognition task as the *only* measure of statistical learning performance (Karuza, Farmer, Fine, Smith, & Jaeger, 2014; Siegelman, Bogaerts and Frost, 2017; Siegelman, Bogaerts, Kronenfeld, et al. 2018). In response to these concerns, these groups made the exposure phase self-paced (Karuza et al., 2014; Siegelman, Bogaerts, Kronenfeld et al. 2018) or turned this phase into a target detection task (Qi, Sanchez, Georgan, Gabrieli, & Arciuli, 2019) such that response times (RTs) can serve as an additional, and online, index of VSL. In the self-paced familiarization phase designs, learners show a predictability advantage such that their RTs to predictable elements (e.g., the second element and third element of the triplets) are faster than their RTs to less predictable elements (e.g., the first element of a triplet). Siegelman, Bogaerts, Kronenfeld et al., detected such predictability advantage using a self-paced VSL in adults. Van Witteloostuijn, Lammertink, Boersma, Wijnen and Rispens (2019) detected it using a self-paced VSL task in children aged between five and eight years old. In the target detection task, learning of the triplets is observed as learners (both children and adults) become faster at detecting the target (which is always the third element of a triplet, and thus predictable if one is sensitive to the triplet structure) over time (see Qi et al., 2019). Finally, Siegelman, Bogaerts and Frost (2017) also gave recommendations on how to expand the offline test phase with different types of test items. Van Witteloostuijn, Lammertink et al. (2019) implemented both the recommended online measure and offline measure in a child-friendly version of the task, and we use their task in the present study.

5.1.5 The present study

The aim of the present study is thus to investigate whether children with DLD have a domain-general statistical learning deficit. In doing so, we compare VSL performance between children with DLD and their typically developing peers, using a self-paced online measure of learning (Siegelman, Bogaerts, Kronenfeld et al., 2018; van Witteloostuijn, Lammertink et al., 2019) and two offline measures of learning (Siegelman, Bogaerts, & Frost, 2017). Our first research question is whether children with DLD have a nonverbal VSL deficit. We expect to observe such a deficit, since we hypothesize that a domain-general statistical learning deficit underlies the language problems in these children. Our second research question concerns the putative association between VSL and literacy in DLD. As van der Kleij et al. (2018) report that growth in pseudoword reading, but not word reading, is associated with serial reaction time performance in children with dyslexia, we will explore the correlations between VSL and reading words and reading pseudowords separately.

5.2 Method

5.2.1 Participants

The present study is part of a larger research project on the relation between statistical learning, grammar and literacy acquisition in children (see Procedure), and consequently our sample of participants overlaps with those reported on in other studies with different research questions (Lammertink, Boersma, Wijnen, & Rispens, 2019, under review [Chapter 4 and Chapter 6 of this dissertation respectively); van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019a, 2019b, submitted).

The two groups of children that participated in the present study – children with DLD and typically developing children – are matched on gender, age (maximal difference of three months), nonverbal intelligence and socioeconomic status (SES). A combined score that takes the average education level, average income and average working status of the people living in a particular district (defined per zip code) is used as a proxy for SES (Sociaal Cultureel Planbureau, 2016). The score has been designed to have a Dutch average of 0 and higher scores indicate higher SES. SES estimates for the children with DLD are based on either their home address (*N* = 22) or school address (*N* = 14). SES estimates for the typically developing children are based on their school address (four different schools across the Netherlands). Ethical approval for this study was obtained from the ethical review committee of the University of Amsterdam, Faculty of Humanities. For the children with DLD, informed consent was given by the children's parents or caregivers prior to participating in the study. Typically developing children were enrolled on an opt-out basis.

Children with DLD. As also described in Lammertink, Boersma, Wijnen and Rispens (2019 [Chapter 4 of this dissertation]) and Lammertink, Boersma, Wijnen and Rispens (under review [Chapter 6 of this dissertation]), 37 children with DLD, aged seven to eleven years old, took part in the study. The children with DLD were recruited via four national organizations in the Netherlands, via an association for parents with children with DLD and self-employed speech therapists. Children had to be diagnosed with DLD by a licensed clinician, taking the following criteria into account: (1) a proficiency score 1.5 *SD* 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, (2) they had at least one parent who is a native speaker of Dutch and (3) they had not been diagnosed with Autism Spectrum Disorder, Attention Deficit Hyperactivity Disorder (ADHD), or other (neuro)psychological problems. In addition to these criteria, children had to obtain a percentile score of at least 17% on the Raven Progressive Matrices (RCPM; Raven, Raven, & Court 2003) – a standardized measure of nonverbal intelligence that was administered as part of our own test battery. After testing, we had to exclude one child with DLD as it turned out that this child had hearing problems in addition to the diagnosis of DLD. This left us with a sample of 36 children with DLD (8 female, 28 male, M_{age} = 9;1. Age range = 7;8 – 10;4). At the start of the project, we contacted different professionals working with children with DLD in the Netherlands (see above). We informed all the professionals who were involved in the recruitment process that recruitment and testing had to take place within a predetermined testing period that ran from January 2017 to March 2018. We tested as many children as possible in this period. The widths of the confidence intervals for our confirmatory and exploratory research questions will tell us whether the power of the experiment was sufficient to detect a medium-sized effect size. As the number of participants per group (*N* = 36) is relatively large for this type of study (see Discussion), we expect that this should not be a problem.

Typically developing children. Fifty-nine typically developing children, aged seven to eleven years, also took part in the study. The typically developing children were recruited via four different primary schools across the Netherlands. Five of the 59 typically developing children that participated were excluded because their nonverbal intelligence score was lower than 17% and/or because they scored below the normal range (norm score < 8; percentile score < 17) on at least two of the following language tasks: one-minute word reading test (Brus & Voeten, 1979), two-minute nonce-word reading test (Klepel; van den Bos, Spelberg, Scheepstra, & de Vries, 1994), spelling (Schoolvaardigheidstoets spelling; Braams & de Vos, 2015) or sentence recall (CELF-4-NL; Semel, Wiig,

& Secord, 2010). Additionally, one typically developing child was excluded, because this child reported having been diagnosed with ADHD. From the remaining 53 children, we selected 36 children (9 female, 27 male, $\mathrm{M}_{\mathrm{age}}$ = 9;1. Age range = 7;8 – 10;4) that matched best with our DLD sample, taking age, gender, SES and nonverbal intelligence into account. For a summary of the group characteristics, see Table 5.1.

5.2.2 Visual statistical learning task

The VSL task used in the present study is also described in van Witteloostijn, Lammertink et. al., 2019 and modelled after previous studies (Arciuli & Simpson, 2012; Siegelman, Bogaerts, & Frost, 2017; Siegelman, Bogaerts, Kronenfeld et al., 2018). The present VSL task differs from the one described by van Witteloostuijn, Lammertink et al., 2019 on four points: (1) we made the task instructions more explicit (see appendix A5.1); (2) There were two sets of alien triplets, instead of one; (3) All children performed a cover task, and this cover task is different from the one described in van Witteloostuijn, Lammertink et al., 2019; (4) In the offline test phase, the order of tasks was reversed: the triplet completion task was first, the triplet recognition task second.

Online familiarization phase. At the start of the experiment, we told children that they were going to play a game in which they would send aliens off to a spaceship (appendix A5.1). The aliens appeared on the screen, one-by-one, and were sent into the spaceship by pressing the space bar. Every time the child pressed the space bar, the current alien disappeared and the next alien appeared. Each alien was part of a triplet of three aliens that always occurred in the same order (thus in the triplet *ABC, B* always followed *A* and *C* always followed *B*). There were four such triplets (*ABC, DEF, GHI, JKL,* see appendix A5.2). Children were not informed about these triplets, but they were told that some of the aliens really liked each other and therefore stood together in line. Children were asked to watch each alien closely and to try and figure out which aliens belonged together. Each triplet occurred 24 times in the familiarization phase, divided over four blocks of six repetitions of each triplet. Between every two successive blocks, there was a small break in which children were awarded a sticker. The predictability of appearance of individual aliens was dependent on the position of the alien within the triplet: the appearance of the second and third aliens is fully predictable from the appearance of the preceding alien(s) (TP = 1.0). The transitional probability when crossing a triplet boundary, thus going from the third alien to the first element of another triplet is lower (each first alien can be preceded by the third alien from either of the two other triplets), making the appearance of each first alien less predictable (Figure 5.1, Figure adapted from van Witteloostuijn, Lammertink et al., 2019, p.5). There were two constraints on the order of appearance of the triplets: (1) the same triplet never appeared twice in a row (e.g., *ABC, ABC*), and (2) repetitions of pairs of triplets (e.g., *ABC, JKL, ABC, JKL*) were ruled out.

There were two experiment versions that differed with respect to which set of individual aliens comprised a triplet (Appendix A5.2). In each experiment version, there were two randomized orders. We decided to work with two experiment versions and two randomized orders to control for any potential effects of single stimuli, triplets or order of appearance. Finally, the familiarization phase had a cover task: children were instructed that occasionally the exact same alien appeared twice in a row. If this happened, the child had to touch the repeated alien with his/her finger on the screen. In each block, such a repetition occurred three times (e.g., *AABC*, *DEEF* and *GHII*) and we ensured that every individual alien was repeated once over the complete course of the familiarization phase.

Offline test phase. The offline test phase consisted of 40 trials (16 triplet completion trials and 24 triplet recognition trials) to test children's knowledge of the triplets that they were familiarized with (the "base triplets"). The base triplets were contrasted with "foil triplets": four triplets that were created from the same set of twelve aliens, but had never appeared as a triplet during the familiarization phase. We tested children's knowledge of complete base triplets (e.g., *ABC;* triplet completion: *N* = 8*;* triplet recognition: *N* = 8) as well as their knowledge of "base pairs" from within the base triplets (e.g., *AB, BC;* triplet completion: *N* = 8; triplet recognition: *N* = 16; Figure 5.2; Appendix A5.3). In the triplet completion trials, children either completed the missing alien in a base triplet or base pair. The correct answer was always one out of three aliens (three-alternative forced choice task). In the triplet recognition items children were presented with either two complete triplets (the base triplet and one foil triplet: e.g., *ABC* versus *DHL*) or two pairs (a base pair and a foil pair: e.g., *AB* versus *DH*) and we asked the children to pick the triplet or pair that appeared most familiar to them (two-alternative forced choice). In both the triplet completion and triplet recognition trials, we controlled for the position of the correct alien on the screen and for the frequency of foil triplets, pairs and single aliens to avoid continued learning during the triplet recognition trials (Arciuli & Simpson, 2012).

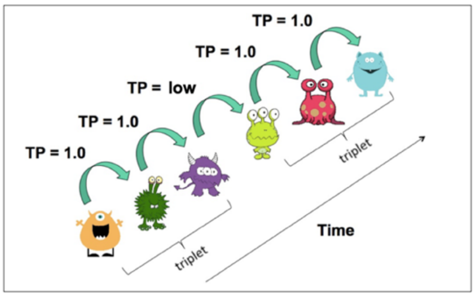


Figure 5.1 The transitional probability (TP) structure in the visual statistical learning task. Note that we adopted this Figure from van Witteloostuijn, Lammertink et al., (2019), p.5

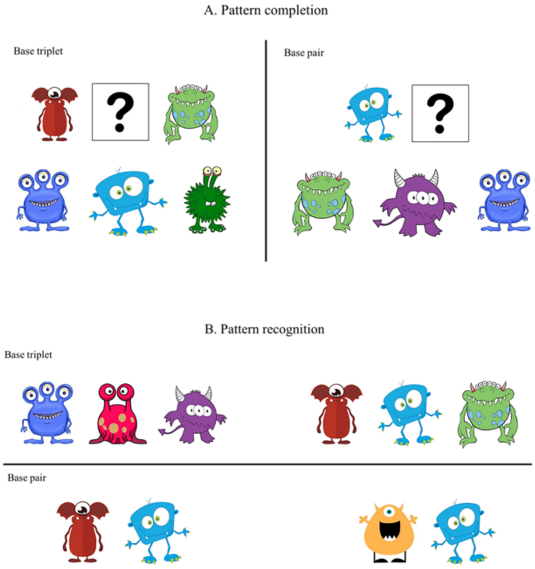


Figure 5.2 (A) Two examples (left: one base triplet; right: one base pair) from the triplet completion trials. Children are asked to replace the question mark with one of the three aliens at the bottom. (B) Two examples (upper row: one base triplet; bottom row: one base pair) from the triplet recognition trials. Children are asked to pick the group of aliens that looks most familiar to them.

Table 5.1 Summary of the group characteristics

	DLD (*N = 36)*	TD (*N* = 36)	Difference DLD – TD		
	Mean [Range]	**Mean [Range]**	***t***	***p***	**95% CI**
Age (years; months)	9;1 [7;8, 10;4]	9;1 [7;8, 10;4]	+0.032	.97	[−0; 3, +0; 3]
Nonverbal intelligence					
Raw	36 [23, 49]	36 [26, 55]	+0.019	.98	[−3, +3]
Standardized (percentiles)	63 [17, 96]	64 [20, 98]			
Socioeconomic status (SES)	+0.22 [−2.57, +2.09]	−0.06 [−1.28, +1.15]	+1.2	.23	[−0.18, +0.75]

(*Table continues*)

Table 5.1 (*Continued*)

	DLD ($N = 36$)	TD ($N = 36$)	Difference DLD – TD		
	Mean [Range]	**Mean [Range]**	t	p	**95% CI**
Expressive morphosyntactic proficiency					
Raw	31 [12, 67]	59 [32, 81]	−9.2	$1.1 \cdot 10^{-13}$	[−35, −22]
Standardized (norm scores)	5[a] [1[a], 13]	11 [3[a], 16]			
Receptive vocabulary knowledge					
Raw	101 [78, 118]	115 [98, 140]	−5.8	$1.6 \cdot 10^{-7}$	[−18, −9]
Standardized (percentiles)	33 [1[a], 84]	63 [6[a], 95]			

Note TD = typically developing; CI = confidence interval. aStandardized 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).

5.2.3 Literacy Tasks

Word reading test. In this task children had one minute to read aloud as many (existing) Dutch words as they could (EMT; Brus, & Voeten, 1979). The raw score was the total number of words read, with a maximum of 116 words. Age-appropriate norm scores were derived from the raw scores. A norm score of 10 corresponds to a percentile score of 50. Norm scores below 8 are interpreted as below average whereas norm scores above 12 are interpreted as above average.

Nonce word reading test: Similarly, as in the word reading task, children were asked to read nonce words aloud. This time, however, they had two minutes to read as many nonce words as they could (Klepel; van den Bos et al., 1994). Again, the maximum number of words to read was 116, and norm scores were derived from the raw scores.

Spelling. In the spelling task (*Schoolvaardigheidstoets spelling*; Braams & de Vos, 2015), the experimenter read aloud a sentence to the child and then instructed the child to write down one word from this sentence. There were 30 items. For each correct written form, children received one point. Age-appropriate percentile scores were derived from the raw scores. Percentile scores below 17 are interpreted as below average whereas scores above 85 are interpreted as above average.

5.2.4 Other cognitive measures

We also took a measure of children's visual spatial short-term memory, their visual spatial working memory and their sustained attention (Table 5.2).

5.2.5 Procedure

As described earlier, the present study is part of a larger research project. The total task battery contained more tasks than reported here. All children that participated in the present study completed the full task battery, and this took two to four sessions (each lasting approximately one hour) per child. The results on the other tasks of our battery, but with the same group children, are reported in Lammertink et al. (2019, [Chapter 4 of this dissertation]) and Lammertink et al. (under review, [Chapter 6 of this dissertation]). Furthermore, a number of the typically developing participants from the same group of 59 typically developing children are also reported on in studies by van Witteloostuijn, Boersma et al. (2019a, 2019b) and van Witteloostuijn et al. (submitted). In her studies, van Witteloostuijn and colleagues uses the performance of the typically developing children to evaluate statistical learning in children with dyslexia.

Table 5.2 Description of the other cognitive measures

Task	**Description**	**Possible range (raw scores)**
Raven's Progressive Matrices (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
Dot Matrix Forward (AMWA; Alloway, 2012)	*Visuospatial short-term memory* Children are presented with a four-by-four matrix in which sequences with dots appeared. Children are asked to point out the position of the dots in the exact same order as presented. The experiment consists of six blocks with each block consisting of maximally six trials. The experiment terminated once a child repeated three or less sequences correct.	0–36
Dot Matrix Backward (AMWA; Alloway, 2012)	*Visuospatial working memory* The task is very similar to the Dot Matrix Backward, with the only difference that children had to point out the position of the dots in reversed order.	0–36
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

5.2.6 Data analysis and hypotheses

We made our data, and the scripts that we used for data analysis available at our Open Science Framework (OSF) page: https://osf.io/8gpjt/.

Online measures of VSL. During the online self-paced familiarization phase, we measured children's RTs to each individual alien (i.e., time between the appearance of the alien on the screen and the child's space bar press) in milliseconds (ms). Prior to analysis, we removed all responses to the three aliens of the first triplet of each block (i.e. four triplets, 12 individual aliens per child). Also, we removed all RTs shorter than 50 ms (DLD; 0.42% of the total observations; typically developing: 0.22% of the total observations). Finally, we normalized the RTs, such that they can be interpreted as optimally distributed z values. These normalized RTs were obtained by first ranking all N raw RT observations, sorting them in increasing order, labelling them with a ranking number r (Baguley, 2012, p. 254-358) and then replacing all observations by the $(r - 0.5) / N$ quantile of the unit normal distribution. We decided a-priori to normalize the raw RTs as with this procedure, we take the data closer to satisfying the assumption of normally distributed model residuals, which is a central assumption of linear mixed-effects model analysis (package *lme4;* Version 1.1.17, Bates, Maechler, Bolker, & Walker 2015; *R* programming language: R Core Team, 2018). Furthermore, the advantage of working with transformed RT data (in general) is that one can include all observations and thus not have to apply an arbitrary criterion in removing *outlier* observations (Simmons, Nelson, & Simonsohn, 2011). As a sanity check we visually inspected the model residuals from the raw RT model and normalized RT model and indeed observe that the residuals of the model with normalized RTs are more symmetrically distributed than the residuals of the model with raw RTs (see histograms on our OSF page: https://osf.io/8gpjt/).

The normalized RTs were analysed using a linear mixed-effects model that fitted normalized RT as a function of the ternary within-subject predictor Predictability (alien 1, alien 2, alien 3), the binary between-subjects predictors Group (DLD, typically developing), TripletVersion (triplets A, triplets B) and TripletOrder (order 1, order 2), and the continuous within-subject predictor Time (repetition of triplets, originally ranging from 1 to 24, after centering and scaling ranging from -1.68 to $+1.65$). All predictors were included in interaction with each other, and the random-effects structure of the model contained by-subject (N = 72) and by-item (N = 12; individual alien) random intercepts and by-subject random slopes for the main effects of Predictability and Time and for their interaction. If children are sensitive to the TPs, then their RTs to predictable aliens (alien 2 and 3) should be faster than their RTs to unpredictable aliens (alien 1). We will refer to this as the “predictability advantage”. The size of the predictability advantage is estimated by the first contrast of the predictor Predictability (with alien 1 coded as $-\frac{2}{3}$ and both alien 2 and alien 3 coded as $+\frac{1}{3}$). A difference in learning between children with DLD and typically developing children may be observed in two ways: either we observe a difference in the average predictability advantage (interaction between Predictability and Group) or in the emergence of a difference in predictability advantage over time (interaction between Time, Predictability and Group). The predictor Group is coded with DLD as $-\frac{1}{2}$ and typically developing coded as $+\frac{1}{2}$. Finally, we included the predictors TripletVersion (coded as $-\frac{1}{2}$ and $+\frac{1}{2}$) and TripletOrder (coded as $-\frac{1}{2}$ and $+\frac{1}{2}$) as they potentially influence learning. These predictors were not of interest to our research question.

Statistical significance of the predictors that estimate the difference in size of the predictability advantage between children with DLD and typically developing children (online measure 1), and the difference in the effect of time on the predictability advantage between both groups of children (online measure 2; i.e., our confirmatory predictors) is assessed via 98.75% profile confidence intervals. These confidence intervals are Bonferroni corrected for multiple testing as we assess the VSL difference with a total of four measures: two online measures and two offline measures.

Offline measures of VSL. Responses in the triplet recognition task and triplet completion task were coded as 1 (correct) or 0 (incorrect), with a maximum score of 24 on the triplet recognition task and a maximum of 16 on the triplet completion task. If children are sensitive to the TPs between the elements, then their correctness probabilities on the offline tasks should exceed chance level (33,3% and 50% respectively). The offline accuracy scores were analysed using generalized linear mixed-effects models (package *lme4,* Bates et al., 2015). For both offline tasks, correctness probability was fitted as a function of the binary predictors Group, TripletVersion and TripletOrder. All predictors were added in interaction with each other and the random effects structure of the model contained a by-subject ($N = 72$) random intercept. We will conclude that children with DLD have a visual statistical learning deficit if their correctness probabilities are significantly lower than those of our typically developing children (main effect of Group). Statistical significance of the confirmatory predictors is assessed via 98.75% profile confidence intervals.

5.3 Results

5.3.1 Background measures

Table 5.3 presents the raw scores and – when available – the standardized scores on the cognitive and literacy tasks for both groups of children. Between-group *t* tests show that children with DLD have lower (raw) scores on all three literacy tasks: word reading ($t(70) = -8.60$, $p = 1.6 \cdot 10^{-12}$); pseudoword reading ($t(70) = -9.34$, $p = 8.7 \cdot 10^{-14}$); and spelling ($t(70) = -12.45$, $p = 5.0 \cdot 10^{-19}$). With a norm score > 7 being interpreted as "average" performance, we observe that 42% of the children with DLD can be classified as "average" readers (i.e., they score > 7 on both the [nonce]word reading tests). For the spelling task, 31% percent of the children with DLD had a percentile score of 17% or higher, indicating that they may be classified as "average" spellers. Finally, we have no evidence that the children with DLD perform differently from typically developing children on the tasks that measured visuospatial short-term memory ($t(69)= -1.83$, $p = .072$), visuospatial working memory ($t(69) = -1.02$, $p = .31$) and sustained attention ($t(70) = -0.78$, $p = .44$). Therefore, we decided not to control for these measures when comparing VSL in children with DLD and typically developing children. Please note that we have missing data on the visuospatial short-term memory and visuospatial working memory for one child with DLD.

Table 5.3 Children,s raw and – when available – norm scores or percentiles for the background measures and group comparisons

	DLD ($N = 36$)	TD ($N = 36$)	Difference DLD – TD		
	Mean [Range]	**Mean [Range]**	***t***	***P***	**95% CI**
Visuospatial Short-term memory					
Raw	22 [13, 30]	24 [13,31]	−1.8	.072	[−4, +0.2]
Visuospatial working memory					
Raw	22 [9, 35]	23 [11, 31]	−1.0	.31	[−4, +1]
Sustained attention					
Raw	7 [1, 10]	8 [3, 10]	−0.78	.44	[−1, +1]
Standardized (norm scores)	8 [1[a], 13]	9 [3[a], 13]			
Word reading					
Raw	33 [5, 69]	63 [31, 87]	−8.6	$1.6 \cdot 10^{-12}$	[−36, −23]

(*Table continues*)

Table 5.3 (*Continued*)

	DLD ($N = 36$)	TD ($N = 36$)	Difference DLD – TD		
	Mean [Range]	Mean [Range]	***t***	***p***	**95% CI**
Word reading					
Standardized (norm scores)	5[a] [1[a], 11]	11 [3[a], 15]	−7.3		
Nonce word reading					
Raw	23 [3, 62]	56 [27, 82]	−9.3	$8.7 \cdot 10^{-14}$	[−40, −26]
Standardized (norm scores)	6[a] [1[a], 11]	11 [7[a], 14]			
Spelling					
Raw	7 [0, 18]	20 [13, 27]	−12	$5.0 \cdot 10^{-19}$	[−15, −11]
Standardized (percentiles)	13[a] [0[a], 59]	54 [19, 94]			

Note TD = Typically developing. CI = confidence interval. aStandardized 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).

5.3.2 Visual statistical learning in DLD

In the sections that present the results of our confirmatory research question (online and offline visual statistical learning) we only present the model estimates for the predictors that are relevant for our hypothesis testing or data checks. The full model outcomes are available on our OSF page: https://osf.io/8gpjt/

Descriptives I. Children's mean RTs to all three alien positions (alien 1, alien 2, alien 3) across the 24 repetitions of each triplet are visualized for the children with DLD and the typically developing children separately in Figure 5.3. Descriptively and pooled over the 24 repetitions, children with DLD respond fastest to the second alien (*M* = 807 ms, *SD* = 624 ms), followed by the third alien (*M* = 812 ms, *SD* = 588 ms), followed by the first alien (*M* = 819 ms, *SD* = 611 ms). Typically developing children respond fastest to the second alien (*M* = 858 ms, *SD* = 555 ms), followed by the first alien (*M* = 859 ms, *SD* = 555 ms), followed by the third alien (*M* = 864 ms, *SD* = 554 ms).

Confirmatory results I: Online measures of VSL. If children are sensitive to the TPs in the VSL task, we expect to observe a predictability advantage. The model estimated that, pooled over the groups, the children responded faster to predictable than to unpredictable aliens (main effect of Predictability: $\Delta z = -0.011$), but this estimate was not significantly different from zero ($t = -0.95$, 98.75% profile CI [−0.041, +0.019], $p = .34$; Table 5.4). The two-way interaction between Predictability and Group estimated that the predictability advantage was larger in our children with DLD than in our typically developing children ($\Delta\Delta z = +0.020$), but this estimate was not significantly different from zero ($t = +0.96$, 98.75% profile CI [−0.032, +0.072], $p = .34$; Table 5.4). To obtain an estimate of the maximal standardized effect size (i.e., the maximal standardized difference between children with and without DLD), we divided the maximal absolute raw effect size (i.e., the greater absolute bound of the confidence interval) by the residual standard deviation of the model (residual *SD* = 0.68). The estimate of the maximal standardized effect size is 0.11 (0.072/0.68). This effect size can be interpreted as a Cohen's *d* effect size (Cohen, 1988) and as it is <0.20, it means that if a difference between children with and without DLD exists at all, the difference will be small.

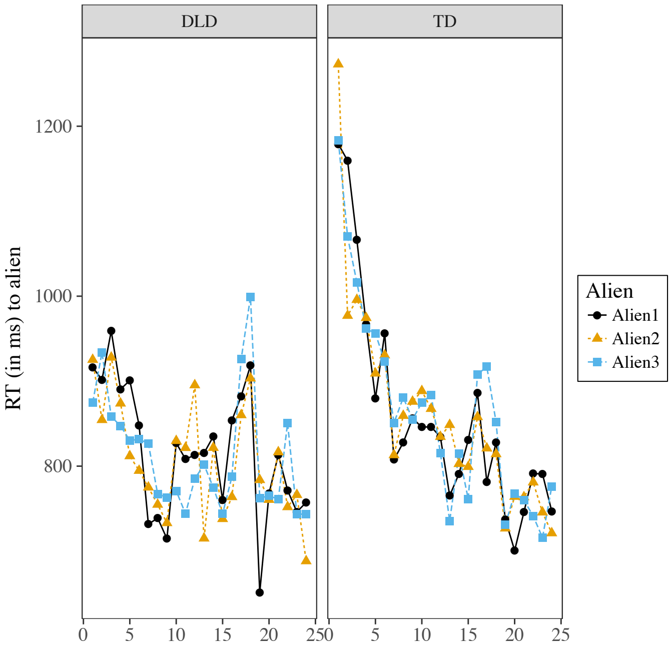


Figure 5.3 Visualization (descriptive) of children's raw (i.e. unnormalized) mean RTs (in ms) to the aliens in first position (black circles), in second position (orange triangles) and third position (blue squares). The left graph shows the RTs of children with DLD, the right graph shows the RTs of typically developing (TD) children. Please note that these raw RTs are only displayed for ease of exposition and that they do not represent the outcome of our confirmatory hypothesis testing. Therefore, (descriptive) differences in these raw RTs cannot be used to interpret the strength of the effects reported later in this paper.

We also looked at the model estimate of children's predictability advantage unfolding over time (interaction between Predictability and Time). Unexpectedly, the model estimated that the predictability advantage decreased over time ($\Delta\Delta z = +0.011$). This decrease was larger for the children with DLD than for the typically developing children ($\Delta\Delta\Delta z = -0.015$). Both the two-way interaction between Predictability and Time and the three-way interaction between Predictability, Time and Group were not significantly different from zero, however (two-way interaction: $t = +1.1$, 98.75% profile CI [−0.014, +0.037], $p = .26$; three-way interaction: $t = -0.73$, 98.75% profile CI [−0.067,

+0.037], *p* = .46; Table 5.4). The estimate of the maximal standardized effect size for a difference in the emergence of a predictability advantage over time between children with DLD and typically developing children is 0.098 (0.067/0.68). Again, the maximal standardized effect size is <0.20 and thus, if a difference between children with and without DLD exists, the difference will be small.

Taken together, the online measures of VSL provide no evidence that children are sensitive or insensitive to the TPs or that sensitivity to the TPs emerges or does not emerge over time. Also, we have no evidence for or against a difference between children with and without DLD.

Confirmatory results II: Offline measures of VSL. For both offline tasks (triplet completion and triplet recognition), the criterion for learning was that the correctness probabilities (i.e., model intercepts) exceed chance level (0.333 for triplet completion and 0.50 for triplet recognition). The intercepts for both offline models estimated that, pooled over both groups of children, children picked the correct answer more than one would expect on the basis of chance (triplet completion: log odds = −0.099, odds = 0.91, probability = 48%; triplet recognition: log odds = +0.53; odds = 1.7, probability = 63%). Both estimates are statistically significantly different from chance probability (triplet completion: *p* = $5.9 \cdot 10^{-7}$, 98.75% CI [41%, 55%]; triplet recognition: *p* = $3.9 \cdot 10^{-7}$; 98.75% CI [57%, 69%]).

If children with DLD learn fewer triplets than the typically developing children, then their correctness probabilities on both tasks should be lower than those of the typically developing children. Indeed, on the triplet completion task, the model estimated that the ratio by which children picked the correct missing alien was 1.1 higher in the typically developing children than in the children with DLD. This odds ratio was not significantly different from 1, however (*z* = +0.66; *p* = .51; 98.75% CI odds ratio [0.67, 2.0], Figure 5.4, Table 5.5A).

For the triplet recognition task, the model estimated that the ratio by which children picked the correct group of aliens was 0.88 times higher (and thus 1.1 times worse) in the typically developing children than in children with DLD. This odds ratio was not significantly different from 1, however (*z* = −0.66; *p* = .51; 98.75% CI odds ratio [0.54, 1.5], Figure 5.4, Table 5.5B).

To check whether the groups separately showed correctness probabilities that exceed chance expectations, we fitted two additional models for both tasks in which we re-referenced the contrast coding for the predictor of Group such that we obtained estimates for the children with DLD (with contrasts set as DLD 0;

typically developing +1) and the typically developing children (DLD +1; TD: 0) separately. For both groups of children, and for both tasks, the estimates were significantly different from chance (Table 5.5A, Table 5.5B).

Taken together, for both populations of children, and for both tasks we conclude that children can learn which aliens belong together. We have no evidence for or against a difference between DLD and typically developing either the completion task or the recognition task.

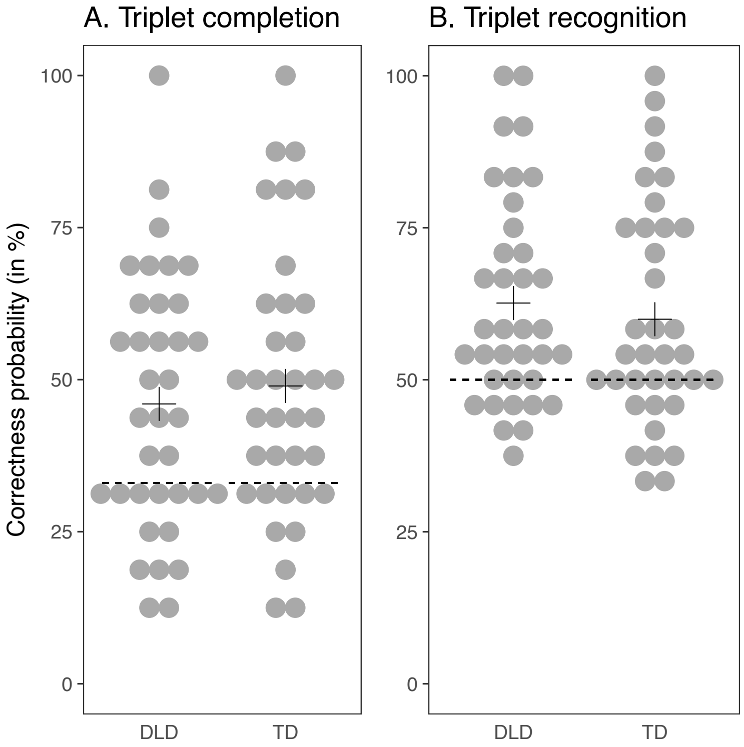


Figure 5.4 Children's individual correctness probabilities on the triplet completion task (left) and triplet recognition task (right). The dashed lines represent chance probability (33.3% for the triplet completion task and 50% for the triplet recognition task). The crosses indicate the mean correctness probabilities per group (DLD and typically developing [TD]). Please note that we did not obtain these correctness probabilities from the statistical model. These descriptive data are only displayed for ease of exposition and do not represent the outcome of the generalized linear mixed model. Therefore, (descriptive) differences in this plot cannot be used to interpret the strength of the effects reported later in this paper.

Table 5.4 Outcome of the relevant estimates from the linear mixed-effects model (normalized response times; 19807 observations)

***Random effects of subject (N* = 72)**		**Correlation**				
	SD (Δ z)	Intercept	Time	Predictability	Alien2 vs. Alien3	Time × Predictability
Intercept	0.65					
Time	0.17	+.27				
Predictability	0.0099	−.98	−.45			
Alien2 vs. Alien3	0.015	+.10	+.28	−.15		
Time × Predictability	0.0083	−.81	−.09	+.77	−.59	
Time × Alien2 vs. Alien3	0.032	−.40	+.32	+.31	+.85	−.07

(*Table continues*)

Table 5.4 (*Continued*)

Random effects of alien (*N* =12)	
	SD(Δ *z*)
Intercept	0.045
Residual	0.68

Fixed effect	***β* (Δ *z*)**	**98.75% confidence interval (Δ *z*)**	***t***	***p***
Predictability × Group[a]	+0.020	[−0.032, +0.072]	+0.96	0.34
Time x Predictability × Group[a]	−0.015	[−0.067, +0.036]	−0.73	0.46
Predictability[b]	−0.011	[−0.038, +0.016]	−0.95	0.34
Time × Predictability[b]	+0.011	[−0.012, +0.035]	+1.1	0.26

Note. The full model outcome (including all estimates) can be viewed in the Rmarkdown at: https://osf.io/8gpjt/.
[a]Relevance is confirmatory. [b]Relevance is data check.

Table 5.5 Outcomes of the relevant estimates for (A) the triplet completion task and (B) triplet recognition task (generalized linear mixed-effects models on correctness probabilities)

(A) Triplet completion task (1152 observations)						
Random effects of subject ($N = 72$)						***SD* (log odds)**
Intercept						0.72
Fixed effect	**β_{model} (log odds)**	**$\beta_{\text{transf.}}$ (odds)**	**$\beta_{\text{transf.}}$ (prob.)**	**98.75% CI (odds)**	**98.75% CI (probability)**	***p***
Group[a]	+0.14	1.1	-	[0.67, 2.0]	-	0.51
Intercept[b]	−0.099	0.91	48%	[0.71, 1.2]	[41%,55%]	$5.9\cdot10^{-7}$
Intercept DLD[b]	−0.17	0.84	46%	[0.59, 1.2]	[36%,56%]	$9.9\cdot10^{-4}$
Intercept TD[b]	−0.029	0.97	49%	[0.69, 1.4]	[41%,59%]	$3.9\cdot10^{-5}$

(*Table continues*)

Table 5.5 (*Continued*)

(B) Triplet recognition task (1728 observations)						
Random effects of subject ($N = 72$)						***SD* (log odds)**
Intercept						0.68
Fixed effect	**β_{model} (log odds)**	**$\beta_{\text{transf.}}$ (odds)**	**$\beta_{\text{transf.}}$ (prob.)**	**98.75% CI (odds)**	**98.75% CI (probability)**	***p***
Group[a]	−0.13	0.88	-	[0.54, 1.5]	-	0.51
Intercept[b]	+0.53	1.7	63%	[1.4, 2.1]	[57%,69%]	$3.9 \cdot 10^{-7}$
Intercept DLD[b]	+0.59	1.8	64%	[1.3, 2.6]	[56%,73%]	$3.5 \cdot 10^{-5}$
Intercept TD[b]	+0.47	1.6	61%	[1.2, 2.2]	[53%,69%]	$8.5 \cdot 10^{-4}$

Note. The full model outcome (including all estimates and refitted models) can be viewed in the Rmarkdown script at: https://osf.io/8gpjt/. TD = typically developing. [a]Relevance is confirmatory. [b]Relevance is data check. CI = confidence interval; transf. = transformed; prob. = probability.

Exploratory results: The link between literacy and VSL. To see if there is an association between VSL and literacy in children with DLD, we averaged children's offline VSL measures (triplet completion and triplet recognition), as children's scores on these tasks were positively correlated, and significantly different from zero (Pearson *r* (34) = +.67; 95% CI [+.44, +.82]).

None of the correlations between VSL and literacy were significantly different from zero (word reading: Pearson *r* (34) = +.070, 95% CI [−.26, +.39; pseudoword reading: Pearson *r* (34) = −.014, 95% CI [−.34, +.32]; spelling: Pearson *r* (34) = +.13, 95% CI [−.20, +.44]; Figure 5.5). Although not part of our hypothesis testing, we also explored the correlations between VSL and literacy in the typically developing children. None of the explored correlations in the typically developing children were significantly different from zero (see output at our OSF page: https://osf.io/8gpjt/).

Taken together, we cannot conclude that offline VSL associates (or does not associate) with individual differences in literacy performance in children with DLD.

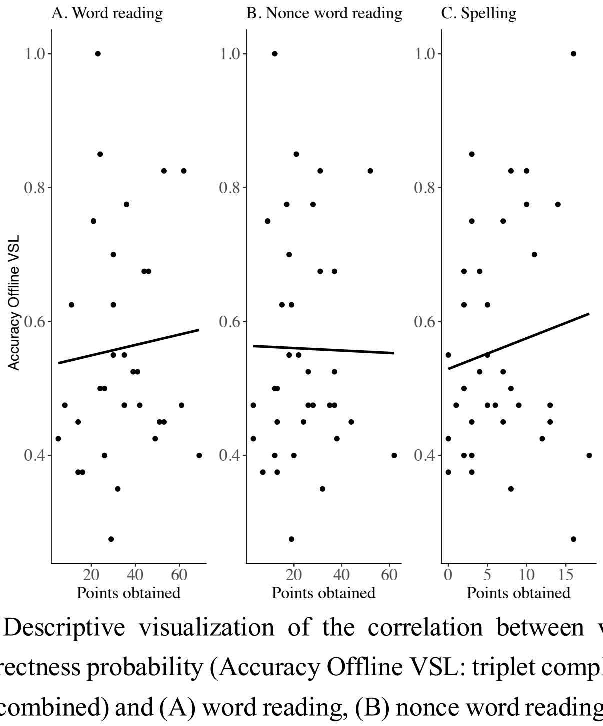


Figure 5.5 Descriptive visualization of the correlation between visual statistical learning correctness probability (Accuracy Offline VSL: triplet completion and triplet recognition combined) and (A) word reading, (B) nonce word reading and (c) spelling in children with DLD.

5.4 Discussion

The main aim of the present study was to assess whether children with DLD have a nonverbal visual statistical learning deficit. We had expected to observe such deficit, since we hypothesized that a domain-general statistical learning deficit underlies the language problems observed in children with DLD. The outcomes of this study provide no evidence for or against such domain-general visual statistical learning deficit in children with DLD. Neither with the online VSL measures nor with the offline VSL measures did we detect a difference in learning between children with and without DLD. Null results, however, can never be used to prove that an effect is absent. Therefore, we can only assign meaning to our findings by showing that, if a difference would exist at all, this difference would be small. We estimated the magnitude of the DLD–typically-developing differences using estimates of the maximal standardized effect sizes (see Results), and found that the maximal standardized effect sizes for both our online measures are below 0.20, meaning that if a DLD–typically-developing difference existed at all, this difference would be small (Cohen, 1988). With the offline measures, we observe that children with DLD either perform maximally 2.0 times worse (upper bound CI) or 1.5 times better (lower bound CI) than typically developing children on the triplet completion task, and maximally 1.5 times worse or 2.0 times better on the triplet recognition task. As there is no general consensus on how to interpret the magnitude of odds ratio effect sizes, we refrain from calling these effect sizes small, medium or large (but see Chen, Cohen, & Chen, 2010).

A limitation of the present study is that the online measures could not detect children's learning of the visual regularities. Therefore, even if a difference between children with and without DLD exists, it is the question whether such a difference will be meaningful. Our small (and statistically nonsignificant) result for the online measure (pooled over groups) of $\Delta z = -0.011$ falls within the (statistically significant) predictability advantage found (for typically developing children) by van Witteloostuijn, Lammertink et al. (2019) which ranged from $\Delta z = -0.114$ to $\Delta z = -0.002$. Van Witteloostuijn, Lammertink et al. (2019) already concluded that the predictability advantage effect can be called small, meaning that if it could be detected at all, the effect may be too small to be reliably detected across studies or between different participant groups. As such the outcomes of the present study fit within a series of recently published papers that investigate the psychometric properties of statistical learning designs. These studies address

(a) the reliability of statistical learning tasks in their ability to capture individual differences in children's (language) learning ability (Arnon, 2019), but also (b) the validity of the tasks in measuring the construct of statistical learning (West et al., 2017). As the present study was not designed to assess the psychometric properties (i.e., split-half reliability and test-retest reliability) of our visual task we cannot draw any conclusions with respect to these issues. Nevertheless, we deem it important to place our study within this debate and to refer the interested reader to relevant papers on this issue (e.g., Arnon, 2019; Siegelman, Bogaerts, & Frost, 2017; West et al., 2017).

Interestingly, our offline measures of VSL indicate that both children with and without DLD are sensitive to the transitional probabilities between the aliens. The children completed and recognized the triplets with correctness probabilities that exceed chance expectation. This may be a preliminary indication that children with DLD are sensitive to TPs in the nonverbal visual domain. This conclusion could not be drawn in Noonan (2018), who also studied VSL in children with DLD, because Noonan could not detect a learning effect in children with and without DLD. It may thus be illuminating to highlight some differences between both studies. Firstly, as we used a self-paced familiarization phase, children were exposed to the stimuli at their own pace. This is different from the Noonan study in which the children were presented with the stimuli at a fixed presentation rate. Secondly, in line with the task instructions given by Siegelman, Bogaerts, Kronenfeld et al., (2018), we instructed the children to pay attention to the order in which the aliens appeared. Even though with these instructions we gave no information about the triplet patterns, our instructions are likely to be more explicit than the "deliberately vague (Noonan, 2018, p.84)" instructions given by Noonan. Thirdly, the stimuli that we used were more colourful, less abstract and thus easier to verbalize than the black, abstract shapes used in Noonan her study. Finally, the present study contained fewer triplets (four triplets) than the study by Noonan (five triplets). We speculate that the abovementioned differences made learning of the structure in present study easier or more explicit than in the study by Noonan. As offline measures of statistical learning are proposed to measure more explicit representations of acquired knowledge (Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015), this may be one of the reasons that we did detect a learning effect in the offline measures.

For the children with DLD, we also investigated the link between their VSL performance and literacy skills, but found no evidence for or against the existence of such a link. The confidence intervals for the association between VSL and our literacy measures ranged from $r = -.26$ to $+.39$ (reading), $r = -.34$ to $+.32$ (pseudoword reading) and from $r = -.20$ to $+.44$ (spelling). The estimated upper bounds of the "standardized" effect sizes for these associations are $R^2 = .15$ ($.39^2$), $R^2 = .10$ ($.32^2$) and $R^2 = .19$ ($.44^2$) respectively, indicating that if associations exist, these may be small in size (as all standardized effect sizes are <0.20 Cohen, 1988). Null results for the relationship between VSL and literacy have recently been reported in other studies with typically developing children (e.g., Schmalz, Moll, Mulatti, & Schulte-Körne, 2018) and in children with DLD (Noonan, 2018).

Given these small effects, the only notable – and probably unsatisfactory – conclusion that we can draw is that the currently available measures of VSL are not sensitive enough to detect differences in VSL between children with DLD and typically developing children (see Arnon, 2019; Arciuli & Conway, 2018; Noonan, 2018; Schmalz et al. 2018, and West et al. 2017, for similar conclusions). We do believe that publication of our null results is important, however. Null results should be published to overcome existing publication biases (van Witteloostuijn et al., 2017; Schmalz et al., 2017) and, because the data should be available for researchers who wish to conduct meta-analyses on this topic.

We have reasons to believe that our null results are not the result of the power of our study being too low to detect the effects under examination: Firstly, in comparison to serial reaction time studies, the number of children with DLD tested for the present study is relatively large (only two out of the eleven published serial reaction time studies tested more than 36 children, Hsu and Bishop, 2014a; Conti-Ramsden, Ullman, & Lum, 2015). Secondly, looking at our outcomes we observe (a) a learning effect with our offline measures of learning, (b) a small DLD–typically-developing difference in online visual statistical learning and (c) small correlations between visual statistical learning and literacy in children with DLD. The detection of an effect (a) indicates that we tested sufficiently children to detect offline visual statistical learning. As for (b) and (c), the confidence intervals of the standardized effect sizes for these effects indicate that if the effects exist, the true effects lie between 0 and small; that's a small range. In an underpowered study this range would have been large. Finally, as we selected our children with DLD according to strict in- and exclusion criteria, we do not think that our results are driven by the use of an unrepresentative group of children with DLD. This claim is supported by our background measures in which children with

DLD show impairments in sentence recall, receptive vocabulary knowledge (clinical markers of the disorder) and reading performance, as compared to their typically developing peers.

At this point we would also like to reiterate that the theoretical question on the domain-generality of the statistical learning deficit is important (Elleman, Steacy, & Compton, 2019; Arciuli and Conway, 2018). Results of the present study provide evidence that children with DLD are sensitive to nonverbal regularities in the visual domain. From this we tentatively conclude that if children with DLD have a statistical learning deficit, this deficit may not be domain-general. Furthermore, in light of the linguistic entrenchment hypothesis as put forward by Siegelman, Bogaerts, Elazar et al. (2018) another possibility is that the statistical learning deficit with linguistic materials observed in children with DLD (for an overview of two meta-analysis supporting this claim see: Lammertink et al., 2017 [Chapter 2 of this dissertation], Obeid et al. 2016) does not necessarily reflects reduced sensitivity to statistical regularities, but that – due to their language deficit – children with DLD have fewer expectations on the underlying structure than typically developing children. Following this line of reasoning, in order to test the hypothesis that children with DLD are less sensitive to statistical regularities, we need to show that it is not their reduced *prior knowledge* of structure that impacts their statistical learning performance. The challenge is thus to develop tasks that are able to detect learning of statistical regularities in verbal and nonverbal materials while controlling for prior knowledge and individual differences of such knowledge.

Acknowledgements

This work is part of the research programme Vernieuwingsimpuls – Vidi (examining the contribution of procedural learning to grammar and literacy in children) with project number 27689005, which is (partly) financed by the Dutch Research Council (NWO), and awarded to Judith Rispens (PI). We extend our gratitude to all children participated and to their parents, teachers and speech therapists who facilitated the children's participation. More specifically we would like to thank Viertaal special education; The Royal Auris Group; The Royal Kentalis Group; Pento and FOSS/ stichting Hoormij for their help in the recruitment of children with DLD. Also, we would like to thank the four primary schools that participated. Finally, we thank Iris Broedelet-Resink, Sascha Couvee, Darlene Keydeniers, Maartje Hoekstra (test assistants), Dirk Jan Vet (technical implementation of the experiment), Noam Siegelman and his colleagues for sharing their thoughts and for sharing the adult version of the VSL experiment and Merel van Witteloostuijn for her extensive help with the design of the experiment.

Chapter 6

Statistical learning in the visuomotor domain and its relation to grammatical proficiency in children with and without developmental language disorder: A conceptual replication and meta-analysis

This chapter is a slightly modified version of the paper that is under review for publication as:

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (under review). Statistical learning in the visuomotor domain and its relation to grammatical proficiency in children with and without DLD: A conceptual replication and meta-analysis.

Data and scripts for analyses: https://osf.io/e9w43/

Abstract

Children with developmental language disorder (DLD) have difficulties acquiring the grammatical rules of their native language. It has been proposed that children's detection of sequential statistical patterns correlates with grammatical proficiency and hence that a deficit in the detection of these regularities may underlie the difficulties with grammar observed in children with DLD. Although there is some empirical evidence supporting this claim, individual studies, both in children with and without DLD, vary in the strength of their reported associations. The aim of the present study is therefore to evaluate the evidence for the proposed association. This is achieved by means of (a) a conceptual replication study on 35 children with DLD and 35 typically developing children who performed the serial reaction time task and a test of grammatical proficiency and (b) a meta-analysis over 18 unique effect sizes, which collectively examined the serial reaction time task – expressive grammar correlation in 139 children with DLD and 453 typically developing children. Both outcomes provide no evidence for or against the existence of the proposed association. Neither do they provide evidence for a difference in the strength of the association between both groups of children.

6.1 Introduction

When acquiring their native language, children unconsciously detect and process structural regularities that facilitate word extraction, the induction of phonological and grammatical categories and the representation of (morpho)syntactic rules (Erickson and Thiessen, 2015; Mintz, 2003; Saffran, Aslin, & Newport, 1996; Wijnen, 2013). It has been proposed that children detect and process these regularities via statistical learning. Evidence that statistical learning may play a role in language development comes from two different sources. Firstly, a number of studies has reported on associations between children's statistical learning ability and different aspects of language (vocabulary size: e.g., Evans, Saffran, & Robbe-Torres, 2009; syntactic processing: Kidd, 2012; Kidd & Arciuli, 2016; Misyak, Christiansen, & Tomblin 2010; Misyak & Christiansen, 2012; grammatical proficiency: Hamrick, Lum, & Ullman, 2018; reading: Arciuli, 2018 and spelling: Treiman, 2018). Secondly, there is evidence for a statistical learning deficit in children who have developmental language disorder (DLD, Evans et al., 2009; Hsu & Bishop, 2014a; Lammertink, Boersma, Wijnen, & Rispens, 2017 [Chapter 2 of this dissertation], 2019 [Chapter 4 of this dissertation]; Obeid, Brooks, Powers, Gillespie-Lynch, & Lum 2016). By definition, children with DLD exhibit difficulties with language, across multiple areas (lexicon, phonotactics, morphology, morphosyntax, syntax, discourse) in the absence of a known biomedical cause, intellectual disability, or unfavourable psycho-social/educational conditions (Bishop, Snowling, Thompson, & Greenhalgh, 2017). Despite heterogeneity in the language difficulties observed across children with DLD, almost all children with DLD struggle with the acquisition of the rule-based aspects (i.e., morphology, syntax, morphosyntax, phonology and phonotactics) of language (Leonard, 2014). Given that the detection of these rule-based aspects of language may depend on the detection of sequential statistical regularities, their problems with these components of language may be explained by a statistical learning deficit (or a procedural learning deficit, see below; Evans et al., 2009; Hsu & Bishop, 2014a; Lammertink et al. 2017 [Chapter 2 of this dissertation], 2019 [Chapter 4 of this dissertation]; Obeid et al., 2016; Ullman & Pierpont, 2005; Ullman & Pullman, 2015).

6.1.1 A deficit in the detection of sequential regularities

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

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

In summary, both the statistical learning deficit hypothesis and the procedural deficit hypothesis argue that children with DLD may have a deficit in their detection of sequential patterns and both accounts predict that typically developing children outperform children with DLD on any learning task that requires the detection of sequential statistical patterns (the serial reaction time task being a prime example of such a task). Evidence for the existence of a sequential learning deficit in children with DLD as compared to typically developing children comes from studies that investigated this type of learning in both groups of children in the auditory domain (see meta-analysis by Lammertink et al., 2017 [Chapter 2 of this dissertation]) and in the visuomotor domain (see meta-analysis by Lum, Conti-Ramsden, Morgan, & Ullman, 2014). Also, the meta-analysis by Obeid et al. (2016) that included studies from both domains, concludes that children with DLD have a statistical learning deficit. Furthermore, both the statistical learning deficit hypothesis and the procedural learning deficit hypothesis predict that children’s performance on the serial reaction time task correlates with grammatical proficiency. A quantitative summary (meta-analysis) of studies investigating such associations in typically developing children learning their first language provided evidence that this is indeed the case (Hamrick et al., 2018). Although the correlation between serial reaction time task performance and grammatical proficiency in children with DLD has been investigated (see next section) in several studies, a qualitative summary of all these studies does not exist yet, but is needed in order to obtain an estimate of the strength of the sequential statistical learning – grammatical proficiency relationship in children with and without DLD. Also, a meta-analysis allows for exploration of moderators of the association that may be difficult to assess with one single study (Black & Bergmann, 2017), for example the effect of age and sequence type (first-order conditional versus second-order conditional, as explained later on).

6.1.2 Statistical learning and grammatical proficiency: The need for replication

The discussion above reveals that there is some empirical evidence that sequential statistical learning (measured with the serial reaction time task) correlates with grammatical proficiency in typically developing children. At the same time, a closer look at the outcome of Hamrick et al.'s meta-analysis reveals that the 95% confidence interval of the average weighted correlation between serial reaction time task proficiency and grammatical proficiency ranges from +.009 to +.495. This means that, in the sense of Cohen (1992), the strength of the association in typically developing children varies between being "small" and being "medium to large". This relative wide confidence interval indicates that the strength of the associations reported in individual studies varies strongly. Indeed, in the studies on typically developing children, the point estimate correlations run from $r = -.43$ (Spit & Rispens, 2018) to +.67 (Kidd, 2012). Also, in studies on this association in children with DLD the observed range of point estimates is large: the point estimates run from $r = -.46$ (Gabriel, Meulemans, Parisse, & Maillart, 2015) to +.46 (Gabriel, Maillart, Stefaniek, Lejeune, Desmottes, & Meulemans, 2013). All together, this suggests a large variability in the size and existence of the proposed association between children's serial reaction time performance and their grammatical proficiency, and thus that the association may not be as robust as commonly thought.

Motivated by these apparent large differences in observed associations, as well as the general replication crisis and the documented existence of publication biases ("file drawer problem") in developmental psychology (Frank et al., 2017), the aim of the present study is to evaluate the existence and strength of the association. This is done by (a) a conceptual replication of previous experiments on a visuomotoric statistical learning deficit in children with DLD and (b) a quantitative summary (meta-analysis) of the studies that investigated the proposed association between serial reaction time performance and grammatical proficiency in children with and without DLD. This meta-analysis also allows us to assess the evidence for publication bias and to explore whether the serial reaction time task–grammatical proficiency correlation differs between children with and without DLD. It may be important to highlight that our meta-analysis serves a different goal than the meta-analysis on serial reaction time performance and grammatical proficiency conducted by Hamrick et al. (2018). Hamrick et al. aimed to test the predictions of the declarative/procedural model in first and second language learners (Ullman, 2014), whereas we focus (a) on the relation between serial reaction time performance and grammatical proficiency only, leaving the relationship between declarative learning and lexical knowledge aside, and (b) we focus on different populations, namely children with and without DLD. This different focus, together with the inclusion of studies on typically developing children that were published after Hamrick et al. (2018) completed their analysis, makes our analysis substantially different from the one conducted by Hamrick et al. (additional studies that we include: Clark & Lum, 2017; Desmottes, Maillart, & Meulemans, 2017; Hani, 2015; Hsu & Bishop, 2014a; Kuppuraj, Rao & Bishop, 2016; Mimeau, Coleman, & Donlan, 2016; Obeid, 2017; Park, Miller, Rosenbaum et al., 2018; Spit & Rispens, 2018; West, Vadillo, Shanks, & Hulme, 2017; West, Shanks, & Hulme, 2018).

6.1.3 The serial reaction time task

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

In the experimental part (i.e., our conceptual replication) of the present study, we use a serial reaction time task identical to the one used by Lum and

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

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

Study 1: Experimental study

6.2 Methods experimental study

6.2.1 Participants

Thirty-seven children with DLD and fifty-nine typically developing children, aged between seven and twelve years of age, participated in the experiment. We informed everyone involved in the recruitment process that recruitment and testing had to fit within a predetermined testing period that ran from January 2017 to March 2018, and we recruited and tested as many children as possible in the available recruitment time. We obtained approval from the ethical review committee of the University of Amsterdam, Faculty of Humanities. For the participants with DLD, their parents or caregivers gave informed consent prior to their children's participation in the study. Typically developing children were enrolled on an opt-out basis. As explained in the Procedure section, the same children with and without DLD also participated in Lammertink, Boersma, Wijnen and Rispens (2020, [Chapter 5 of this dissertation]) and in Lammertink et al. (2019, [Chapter 4 of this dissertation]), but there is no overlap in the tasks. Furthermore, data from a subset of the typically developing children that participated in this study are also reported on in van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019a, 2019b, submitted).

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

Typically developing children. We recruited the typically developing children through four different primary schools across the Netherlands. We used the Raven Progressive Matrices subtest (Raven et al., 2003), the one-minute-real-word reading test (Brus & Voeten, 1979), the two-minute nonce-word reading test (van den Bos, Spelberg, Scheepstra, & de Vries, 1994), a test of spelling (Braams & de Vos, 2015) and the sentence recall test from the Clinical Evaluation of Language Fundamentals–Dutch version (Semel, Wiig, & Secord, 2010) to determine if children met our inclusion criteria for the typically developing children (all these tasks were part of our own task battery, see Procedure section). We excluded children that scored below the normal range on the Raven Progressive Matrices and/or on two or more of the four language tasks mentioned above. Additionally, we also excluded children from the typically developing group if they had been diagnosed with autism spectrum disorder, attention deficit hyperactivity disorder, or with other (neuro)physiological problems. In total, we excluded five children by the first criterion and one child by the second criterion. From the remaining 53 typically developing children, we selected 35 children that matched best with our DLD sample, considering age, gender, socioeconomic status (on the basis of postal code; Sociaal en Cultureel Planbureau, 2017) and nonverbal intelligence (Raven et al., 2003). We refer to Table 6.1 for a summary of the relevant group characteristics.

6.2.2 Materials

Serial reaction time task. We used the serial reaction time task identical to the one used by Lum and Kidd (2012). Children were seated in front of a Microsoft Surface 3 tablet computer screen, with a gamepad controller attached to the computer, which was running E-prime (Version 2.0; 2012) software. A visual stimulus (a cartoon picture of a smiley) appeared repeatedly in one of four marked locations on the screen and we instructed children to press the corresponding button on the gamepad controller as quickly and accurately as possible. Each stimulus was visible until the child pressed the corresponding button, with a maximum response time of 3 seconds. After the child had responded, there was a 250-millisecond interval before the next stimulus appeared. Before the start of the real test, we presented children with 28 practice trials to ensure that they understood the task. Unbeknownst to the children, we had divided the stream of stimuli into seven blocks. The first block (20 trials) and sixth block (60 trials) contained trials in a random sequence ("random trials"), whereas the trials in blocks 2 through 5 and in block 7 followed a 10-item deterministic, first-order conditional sequence that was repeated six times (thus 60 trials per block in total). The sequence, where the numbers 1-4 represent the four locations on the screen, was [4, 2, 3, 1, 2, 4, 3, 1, 4, 3]. We refer to these sequenced blocks as "sequenced blocks" and to block 6 as the "disruption block".

Sentence recall task. We measured children's productivity of (morpho)syntactic rules with the sentence recall task – a subtest of the Dutch Clinical Evaluation of Language Fundamentals test battery (CELF-4-NL; Semel et al., 2010). In this task, children are instructed to recall sentences with increasing length and complexity. Following the guidelines of the CELF-4-NL, responses are assigned points in relation to the number of errors (e.g., omissions, additions, replacements, substitutions, switches, incorrect markings) made in recalling the sentence. Children receive three points for fully correct recalls, two points for recalls with one error, one point for recalls with two or three errors and zero points for recalls with four or more errors, with a maximum number of 93 points. The task terminates when a child scores zero points on five consecutive recalls.

6.2.3 Procedure

All children took part in our larger study on the relation between statistical learning and grammar and literacy proficiency in children. The total task battery contained more tasks than described here (2 additional statistical learning tasks and a set of additional language tasks and cognitive tasks). The other tasks are described in Lammertink et al., (2020, [Chapter 5 of this dissertation]) and Lammertink et al. (2019, [Chapter 4 of this dissertation]).

Table 6.1 Summary of the group characteristics

	DLD ($N = 35$)	TD ($N = 35$)	Difference DLD – TD		
	Mean [Range]	**Mean [Range]**	t	p	**95% CI**
Age (years;months)	9;1 [7;8, 10;4]	9;1 [7;8, 10;4]	+0.028	.98	[−0;3, +0;3]
Nonverbal intelligence					
Raw	36 [23, 49]	36 [26, 55]	+0.24	.81	[−3, +3]
Standardized (percentiles)	64 [17, 96]	62 [20, 98]			
Socioeconomic status (SES)					
Raw	+0.23 [−2.57, +2.09]	−0.030 [−1.28, +1.15]	+1.09	.28	[−0.21, +0.73]
Word reading					
Raw	33 [5, 69]	62 [31, 87]	−8.25	$8.9 \cdot 10^{-12}$	[−36, −21]
Standardized (norm scores)	5[a] [1[a], 11]	11 [3[a], 15]			

(*Table continues*)

Table 6.1 (*Continued*)

	DLD (*N* = 35)	TD (*N* = 35)	Difference DLD – TD		
	Mean [Range]	**Mean [Range]**	***t***	***p***	**95% CI**
Nonce word reading					
Raw	24 [3,62]	56 [27,82]	−8.97	$5.1 \cdot 10^{-13}$	[−39, −25]
Standardized (norm scores)	6^a [1^a, 11]	11 [7^a, 14]			
Spelling					
Raw	7 [0, 18]	20 [13, 27]	−12.1	$2.2 \cdot 10^{-16}$	[−15, −11]
Standardized (Percentiles)	13^a [0^a, 59]	53 [19, 94]			
Expressive grammatical proficiency					
Raw	31 [12, 67]	59 [32, 81]	−9.0	$5.8 \cdot 10^{-13}$	[−34, −22]
Standardized (norm scores)	5^a [1^a, 13]	11 [3^a, 16]			

Note. TD = typically developing; CI = confidence interval. aStandardized 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).

All children completed the full task battery, and this took two to four sessions per child. Each child was tested individually. We randomly allocated each child to one of the six different orders in which task administration took place.

6.2.4 Data analysis

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

In analogy to our earlier work (Lammertink, et al., 2019, 2020 [Chapter 4 and Chapter 5 of this dissertation respectively), we normalized children's raw response times such that they can be interpreted as optimally distributed *z* values (see our analysis script at our Open Science Framework (OSF) page: https://osf.io/e9w43/ and previous work for normalization procedure). Then, we used a linear mixed-effects model that fitted these normalized response times as a function of ternary predictor Block (block 5, block 6, block 7) in interaction with the binary predictor Group (DLD, typically developing children) to assess the size of the statistical learning deficit. The random-effects structure of this model contained by-subject ($N = 70$) and by-position ($N = 4$) random intercepts, by-subject random slopes for the main effect of Block and by-position random slopes for the main effect of Group. The ternary predictor Block was coded such that the first contrast of this predictor ("DisurptionPeak") estimated the size of the disruption peak, with the disruption block coded as $+\frac{2}{3}$ and with both sequenced blocks coded as $-\frac{1}{3}$. This predictor DisruptionPeak can be seen as a sanity check, as finding a positive (and statistically significantly different from zero) estimate means that we detected learning, pooled over both groups of children, in our serial reaction time task. The binary predictor Group was coded with DLD as $-\frac{1}{2}$, and with typically developing children as $+\frac{1}{2}$. A positive (and statistically significantly different from zero) estimate for the interaction between the first contrast of the predictor Block and the predictor Group intends to answer our first confirmatory research question, namely whether children with DLD have smaller disruption peaks than typically developing children. We assessed statistical significance of both estimates via 95% profile confidence intervals and wrote the *get.p.value* function (see Rmarkdown functions script at our OSF) to obtain the corresponding *p* values from the profiles iteratively (see also Lammertink et al., 2019 [Chapter 4 of this dissertation]).

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

6.3 Results experimental study

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

6.3.1 Descriptive results

We have no evidence that children with DLD make more (or fewer) errors than typically developing children (pooled over blocks 2 through 7: accuracy children with DLD = 92%; accuracy TD children = 94%, $t = -0.63$, $p = .53$, 95% CI accuracy group difference [−0.054%, +0.028%]). After removing children's incorrect responses and their responses faster than 50 milliseconds (RT < 50 milliseconds: 0.1% in DLD and 0.07% in typically developing children), we calculated the mean raw response times (in milliseconds) with their corresponding standard deviations (in milliseconds) for each block and each group separately (Table 6.2). These raw response times and standard deviations are computed for ease of exposition only and cannot be used to interpret the strength of effects reported later in this paper.

Table 6.2 Descriptive mean raw response times and standard deviations (in parentheses), both in milliseconds for the sequenced blocks and disruption blocks for the children with DLD and the typically developing children separately

	Block 2 (seq.)	**Block 3 (seq.)**	**Block 4 (seq.)**	**Block 5 (seq.)**	**Block 6 (disr.)**	**Block 7 (seq.)**
DLD	679	685	705	698	784	717
	(327)	(351)	(384)	(399)	(402)	(383)
TD	678	704	700	729	798	708
	(314)	(359)	(354)	(411)	(402)	(357)

Note. TD = typically developing; Seq. = sequenced; disr. = disruption.

6.3.2 Performance on the serial reaction time task

Though not part of our confirmatory hypothesis testing, we did check whether, pooled over both groups of children, we have evidence that children learned the sequence. The predictor that estimated the size of the learning effect ("DisruptionPeak") was positive and statistically significantly different from zero ($\Delta z = +0.25$, $t = 8.18$, 95% profile CI [+0.19, +0.31, $p = 7.4 \cdot 10^{-9}$], from which we conclude that children can learn the sequence. To obtain a standardized effect size of this learning effect we divided the estimate by the residual standard deviation (which is 0.86) of the model. The resulting effect size is 0.29 (0.25/0.86) and can be interpreted as a Cohen's *d* effect size (Cohen, 1988). To answer our first confirmatory research question, we looked at the estimate for the interaction between the predictors DisruptionPeak and Group. This estimate was positive ($\Delta\Delta z = +0.019$): the disruption peak was larger in our typically developing children than in our children with DLD, although the estimate is not significantly different from zero ($t = 0.32$, 95% profile CI [−0.10, +0.14], $p = .75$; effect size = 0.022 [0.019/0.86]). Therefore, we cannot conclude that the size of the disruption peak differs or does not differ between typically developing children and children with DLD in general (Figure 6.1). To further explore whether both groups of children separately showed a statistically significant disruption peak, we fitted two additional models on the exact same data, but with different contrast settings for the predictor Group (e.g., with DLD coded as 0 and TD coded as +1 to estimate the disruption peak in DLD). The estimate for the size of the disruption peaks in both groups of children was positive (DLD: Δz = +0.24; Typically developing: Δz = +0.25) and statistically significantly different from zero (DLD: t = 5.56, 95% profile CI [+0.15, +0.32, $p = 4.6 \cdot 10^{-7}$, point estimate effect size: 0.28; Typically developing: t = 6.02, 95% profile CI [+0.17, +0.34, $p = 6.2 \cdot 10^{-8}$, point estimate effect size: 0.29). From this we conclude that both children with DLD and typically developing children are sensitive to the regularities in the input.

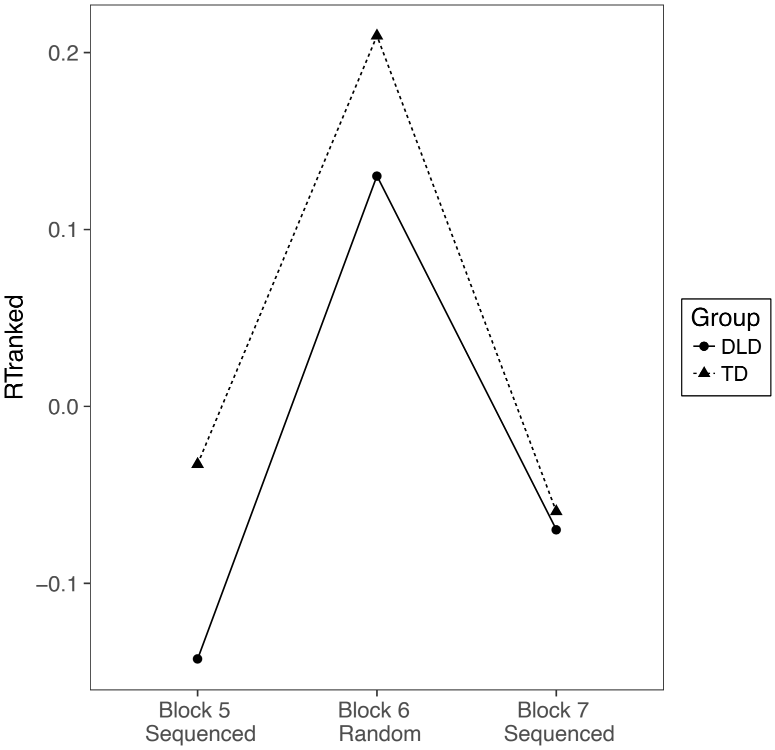


Figure 6.1 Model estimates of the normalized response times to the items across block 5 (sequenced), block 6 (disruption) and block 7 (sequenced). Normalized response times are plotted for the children with DLD (circles, solid line) and typically developing children (triangles, dashed line) separately.

6.3.3 Serial reaction time task performance and expressive grammatical proficiency

To answer our second confirmatory research question, we used the *cor.test* function in R (R core team, 2018) to compute Pearson correlations between the sizes of children's individual disruption peaks and their scores on the sentence recall task. In both groups, the confidence intervals for the correlation include zero and thus we found no evidence for or against a relationship between children's size of the disruption peak and their score on the sentence recall task (DLD: *r* (33) = −.33, 95% CI [−.60, +.00]; TD: *r* (33) = +.18, 95% CI [−.16, +.48], Figure 6.2).

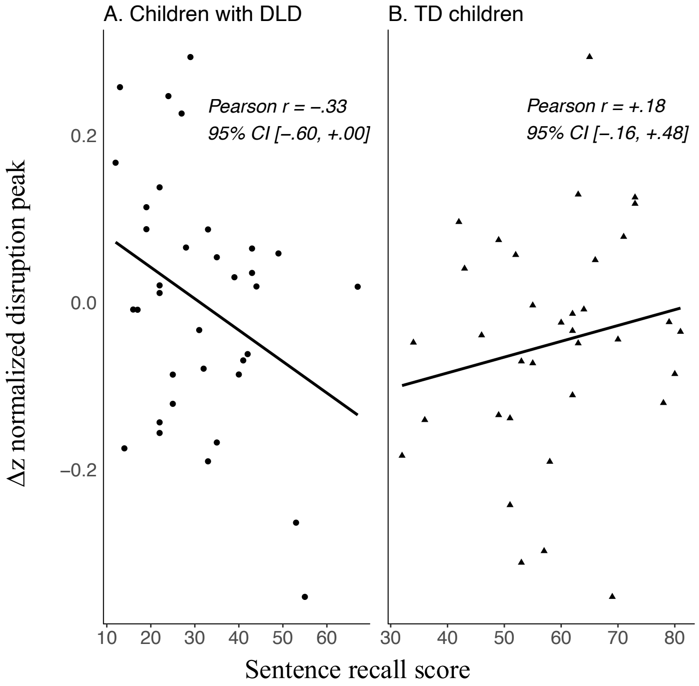


Figure 6.2 Descriptive visualization of the correlation between the size of children's individual disruption peak (centered and scaled, vertical axis) and their average points obtained on the sentence recall task from the CELF (centered and scaled, horizontal axis). The correlation for children with DLD is plotted with black circles and on the left side. The correlation for typically developing children in plotted with black triangles and on the right sight. Each circle and triangle represent the correlation for an individual child. TD = typically developing.

6.4 Discussion experimental study

The experiment was designed to assess the strength of the association between serial reaction time performance and expressive grammar in children with and without DLD. Additionally, we aimed to replicate previous findings showing that children with DLD are less sensitive to structural regularities in the visuomotor domain as compared to their typically developing peers (see meta-analysis Lum et al., 2014). Therefore, we used a serial reaction time task design that is commonly used to assess the difference in performance between children with and without DLD. The task that we used was identical to the one used by Lum and Kidd (2012). Lum and Kidd did not compare serial reaction time task performance between children with and without DLD, but task designs similar to the setup of their task (see Introduction) have been used to assess the presence of a visuomotoric statistical learning deficit in children with DLD (Clark and Lum, 2017; Conti-Ramsden et al. 2015; Hsu & Bishop, 2014a; Park et al., 2018). Unexpectedly, we observed that both groups of children were sensitive to the structural regularities, and we found no evidence for or against a difference in sensitivity between children with and without DLD. To evaluate whether this result is compatible with the standardized effect size for the DLD–typically-developing difference reported in the meta-analysis by Lum et al. (2014), we computed the standardized effect size for our point estimate. This was done by dividing our point estimate (0.019) by the residual standard error of the model (0.86). The resulting standardized effect size (0.022) falls outside the 95% confidence interval of standardized effect sizes reported by Lum et al. (0.071 to 0.584). The difference in point estimate standardized effect size between Lum et al. and the present study is 0.306 (0.328 − 0.022) and as the confidence interval for this difference, which ranges from 0.015 to 0.596, does not include zero, we conclude that our observed effect size is incompatible with the one reported in the meta-analysis by Lum et al. (for a computation of the confidence interval around the point estimate difference, see our OSF page: https://osf.io/e9w43/).

We found no evidence for or against an association between serial reaction time task performance and sentence recall in children with and without DLD. In an attempt to explain these correlational results within the context of previous work on this topic, we noticed that there is no consensus on the existence and strength of the proposed association between serial reaction time performance and expressive grammatical proficiency in children with and without DLD: only a small minority of studies report statistically significant correlations (see Figure 6.5 later in this chapter). Also, the strength of the reported association in children with and without DLD varies across studies (see Introduction). In an attempt to put our result into perspective and to assess the existence of a potential publication bias, we decided to also conduct a meta-analysis on this topic. The meta-analysis is discussed in the following sections. Please note that the focus of this meta-analysis is on the association between serial reaction time performance and expressive grammatical proficiency (rather than receptive grammatical proficiency).

Study 2: Meta-analysis

6.5 Methods meta-analysis

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

6.5.1 Literature search

A first systematic search was conducted by the first author of this paper in February 2018. The search was conducted in five different sources: PubMed, PsycINFO, Education Resources Information Center (ERIC), Linguistics and Language Behavior Abstracts (LLBA) and Open Access Theses and Dissertations (OATD). In addition, the first author also contacted experts in the field (via the LINGUIST List and via the Cogdevsoc list) with requests for access to unpublished data. Altogether, this first search yielded 93 unique articles (91 hits via the databases and 2 hits via the mailing lists; Figure 6.3). A second search in PubMed, PsycINFO and OATD, which served as a reliability check, was done by a research assistant in September 2018. This second search yielded 13 additional potentially relevant unique articles that were not in the output of the first search. Finally, a third search was conducted by another research assistant in January 2019. This third search was conducted as we realized that our initial query focused on studies that included people with DLD/specific language impairment and that therefore, we might have missed articles on serial reaction time task performance in typically developing children. This third search yielded 11 additional potentially relevant unique articles. Thus, in total we screened 115 unique articles on their title and abstract. If, by screening the title and/or abstract, it became clear that the study did not meet the inclusion criteria for the meta-analysis (see Inclusion criteria and study selection), then the study was excluded. For 49 articles or datasets, we read the methods and result sections carefully in order to decide whether or not the study met the inclusion criteria. Eventually, 18 articles (15 published articles, 1 preprint and 2 dissertations) met our inclusion criteria and were included in our database (see Sample description). See our OSF page for Excel spreadsheets with information on why studies were eventually included or excluded for analysis.

6.5.2 Inclusion criteria and study selection

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

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

6.5.3 Sample description

The final sample includes 54 effect sizes pertaining to correlations between an index of serial reaction time performance and grammatical proficiency. Twenty-seven of these 54 effect sizes are correlations with an expressive grammar index. The other 27 effect sizes are correlations with a receptive grammar index. As our research question concerns expressive grammar only, we continue to describe only the dataset that includes correlations between serial reaction time task performance and expressive grammar (but see our OSF page for an exploratory analysis on the correlation with receptive grammar).

After selecting and/or synthesizing effect sizes that came from the same sample of children (see Effect size computation and synthesized effect sizes), the final dataset contained 18 unique correlations between expressive grammar (indexed via a sentence recall or sentence completion task) and serial reaction time task performance (first-order conditional sequence: N = 12; second-order conditional: N = 6) in children with DLD (N = 8 effect sizes, 139 children with DLD) and in typically developing children (N = 10 effect sizes, 453 typically developing children).

6.5.4 Effect size computation

From each study, we extracted the relevant correlation coefficients. If needed, we synthesized effect sizes that came from the same sample of children (see Synthesized effect sizes). The extracted correlations were transformed into Fisher *z* correlations with their corresponding variances (Borenstein 2009, p. 42, Formula A and Formula B in appendix 6.2). All studies, except the study of Hani (2015), reported Pearson *r* correlations. The study by Hani (2015) reported a Kendall's *tau* correlation and therefore we first transformed this correlation into Pearson *r* (Formula C, appendix 6.2) before transforming it into Fisher *z*.

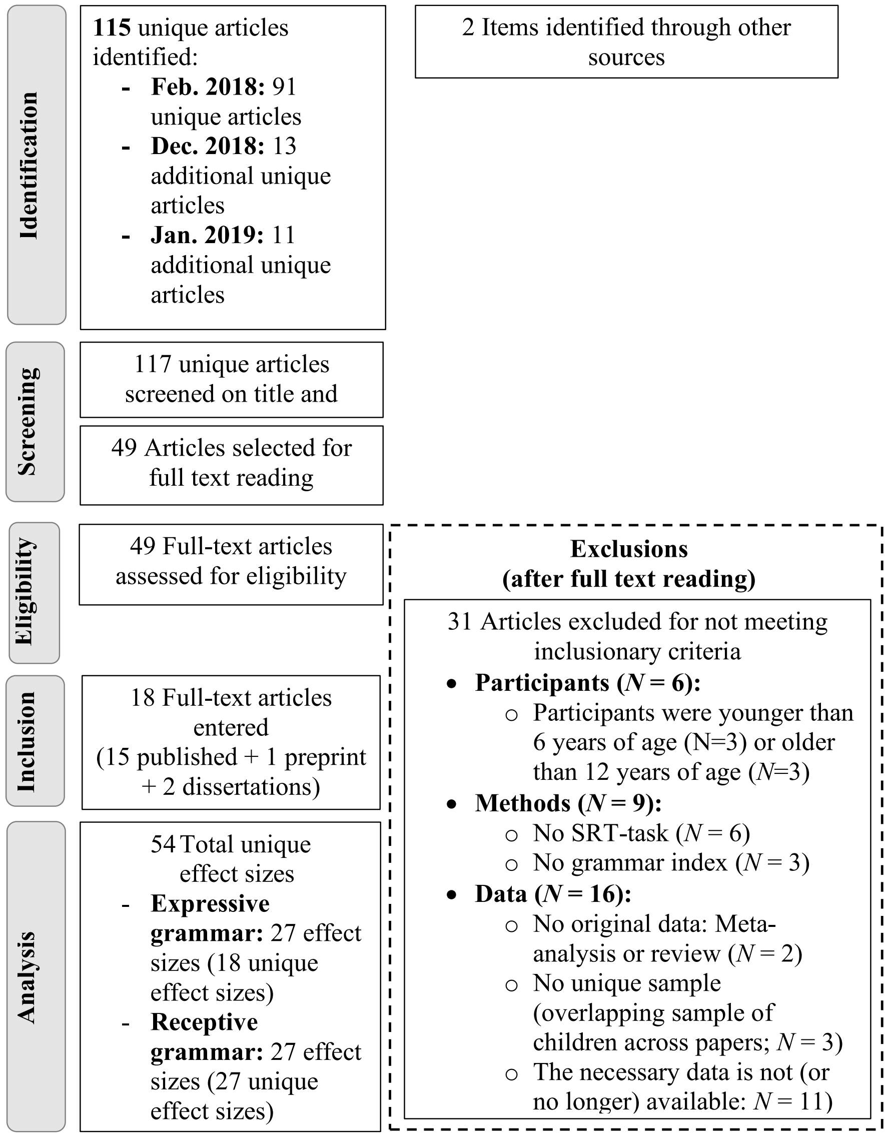

Identification
115 unique articles identified:
- **Feb. 2018:** 91 unique articles
- **Dec. 2018:** 13 additional unique articles
- **Jan. 2019:** 11 additional unique articles
2 Items identified through other sources
Screening
117 unique articles screened on title and
49 Articles selected for full text reading
Eligibility
49 Full-text articles assessed for eligibility
Exclusions (after full text reading)
31 Articles excluded for not meeting inclusionary criteria
Participants (*N* = 6):
Participants were younger than 6 years of age (N=3) or older than 12 years of age (*N*=3)
Methods (*N* = 9):
No SRT-task (*N* = 6)
No grammar index (*N* = 3)
Data (*N* = 16):
No original data: Meta-analysis or review (*N* = 2)
No unique sample (overlapping sample of children across papers; *N* = 3)
The necessary data is not (or no longer) available: *N* = 11)
Inclusion
18 Full-text articles entered (15 published + 1 preprint + 2 dissertations)
Analysis
54 Total unique effect sizes
- **Expressive grammar:** 27 effect sizes (18 unique effect sizes)
- **Receptive grammar:** 27 effect sizes (27 unique effect sizes)

Figure 6.3 Flowchart indicating data exclusion at each stage of the literature search procedure. SRT = serial reaction time task.

6.5.5 Synthesized effect sizes

There were seven articles that reported multiple correlations between serial reaction time task performance and expressive grammatical proficiency in the same group of children. These multiple correlations were reported either because children performed multiple serial reaction time tasks at different timepoints, or because the authors obtained multiple measures of children's expressive grammatical proficiency (e.g., children did both a sentence recall and a sentence formulation task). We cannot include correlations that come from the same group of children in one meta-analysis, as that would violate the assumption of independence. Therefore, we either selected (a) only one of the correlations reported or (b) we computed a synthesized effect size across the multiple correlations reported. We chose option (a) if the multiple correlations were reported for different timepoints, and option (b) if multiple measures of expressive grammar were reported. In the case of solution (a), we decided to select only the correlation reported for the child's first serial reaction time task session (Gabriel, Stefaniek, Maillart, Schmitz, & Meulemans, 2012; Desmottes, Meulemans, & Maillart, 2016a, Desmottes, Maillart et al., 2017; West et al., 2017; 2018)[12]. In the case of solution (b), we computed a synthesized (combined) effect size, and its associated synthesized variance (formulas D and E; Borenstein 2009, p. 227; Desmottes et al., 2016a; Desmottes, Maillart et al., 2017; Park et al. 2018; Obeid, 2017). The resulting synthesized effect sizes are reported in Table 6.3. The calculation of these synthesized effect sizes required knowledge of the correlation between the two measures of expressive grammar. For Park et al. (2018) and for Obeid (2017), we obtained these correlations from the authors. Unfortunately, we did not obtain this information for the Desmottes et al. (2016a) and Desmottes, Maillart, et al. (2017) papers. Therefore, we took these correlations from another paper by the same authors (Desmottes, Meulemans, & Maillart, 2016b) in which they did report the correlations, although for different samples of children.

6.5.6 Data analysis and coding of moderator variables

The main aim of the meta-analysis was to assess the strength of the relationship between serial reaction time task performance and expressive grammatical proficiency in primary-school-aged children. We set out to answer this

[12] Please note that we eventually decided to exclude Desmottes, Meulemans, Patinec, and Maillart (2017), because the authors reported a correlation for the third session only.

confirmatory research with a hierarchical meta-analytic random effects model (see *rma.mv* function from the *metafor* package, Version 2.0.0 in R, Viechtbauer, 2010) in which we fitted the mean weighted correlation as a function of the binary moderator Group, with DLD coded as $-\frac{1}{2}$ and with typically developing coded as $+\frac{1}{2}$. The random-effects structure contained a random intercept for Paper (*N* = 12). Simultaneously we also explored whether the mean weighted correlation is stronger in typically developing children than in children with DLD (as the declarative memory compensation hypothesis may predict; Ullman & Pullman, 2015). A positive (and statistically different from zero) estimate for the predictor Group may be a preliminary indication that this hypothesis is true.

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

Table 6.3 Overview of studies for which we computed a synthesized effect size (i.e. average combined correlation between children' s performance on the serial reaction time task and their scores on the two measures of grammatical proficiency used). Computation of these synthesized effect sizes requires knowledge of the correlation between the two measures of grammatical proficiency. These correlations are reported in the first column

	Correlation (Pearson *r*) between grammar index 1 and grammar index 2		**Synthesized effect size (Pearson *r*) with its corresponding variance (in parentheses)**	
	DLD	**TD**	**DLD**	**TD**
Park et al. (2018)	+.67	+.33	−.24 (+.042)	.00 (+.019)
Obeid et al. (2017)	NA	+.57	NA	−.043 (+.013)
Desmottes, Meuelemans, & Maillart, 2016a	+.82	+.38	−.28 (+.039)	−.073 (+.023)
Desmottes, Maillart, & Meulemans, 2017	+.82	+.38	+.24 (+.047)	+.051 (+.043)

Note. TD = typically developing.

6.7 Results of the Meta-Analysis

6.7.1 Publication bias

To assess the presence of a publication bias in the present meta-analysis, we analysed funnel plot asymmetry (Egger, Smith, Schneider and Minder, 1997) with a linear regression on our funnel plot (Figure 6.4). Visual inspection of our funnel plot suggests that the effect sizes are symmetrically distributed and therefore publication bias seems unlikely. Using the *regtest* function in the *metafor* package (Version 2.0.0) of the statistical programming language R (Viechtbauer, 2010), we found no evidence for or against funnel plot asymmetry (publication bias) in our sample ($z = -1.67$, $p = .096$).

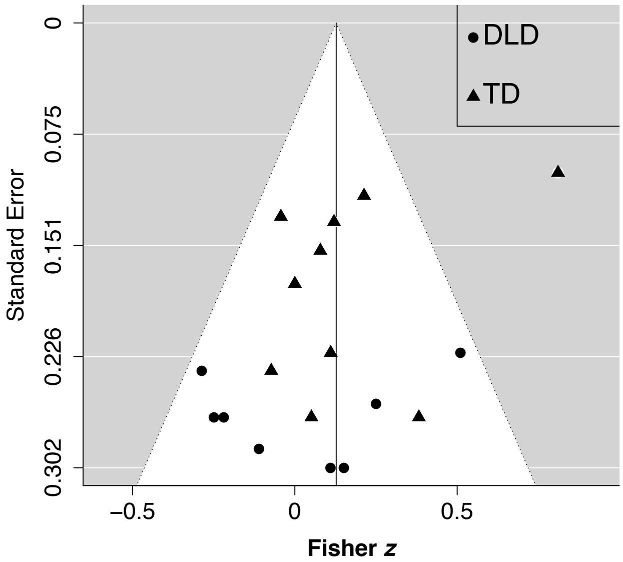


Figure 6.4 Funnel plot showing standard error of the effect size Fisher *z* as a function of the effect size. The vertical line indicates the mean weighted correlation. Dots in black are individual effect sizes from children with DLD, triangles in black represent individual effect sizes from typically developing children. The triangle-shaped unshaded region represents a pseudo confidence interval region with bounds equal to ± 1.96 SE.

6.7.2 Confirmatory meta-Analysis

The model outcome provided no evidence for or against a correlation between serial reaction time task performance and expressive grammatical proficiency in the pooled group of children (Fisher *z* = 0.13, *SE* = 0.084, *z* = 1.51, *p* = .13, 95% CI [−0.038, +0.29], Figure 6.5). For ease of interpretation, we also transformed the Fisher *z* estimate and its 95% confidence interval back into Pearson *r* values: the corresponding values are: *r* = .13, 95% CI [−.038, +.28].

6.7.3 Exploratory analyses

In addition to assessing the strength of the correlation between serial reaction time performance and expressive grammatical proficiency, we also explored whether the strength of this relationship differed between children with and without DLD. The model outcome of the predictor estimated that the strength of the relationship is stronger in typically developing children than in children with DLD. However, the estimate was not significantly different from zero (Fisher *z* = +0.11, *SE* = 0.12, *z* = 0.89, *p* = .37, 95% CI [−0.13, +0.35]; Pearson *r* = +.11, 95% CI [−.13, +.34]).

We also explored whether the mean weighted correlation between serial reaction time performance and expressive grammar, controlled for Group status (DLD versus typically developing) differed as a function of sequence type (first-order condition versus second-order conditional) or age. Model comparisons revealed that we cannot conclude that this is the case. Neither the Group model versus Group–Sequence model comparison (*p* = .26) nor the Group model versus Group–Age model comparison (*p* = .20) was significantly different from zero.

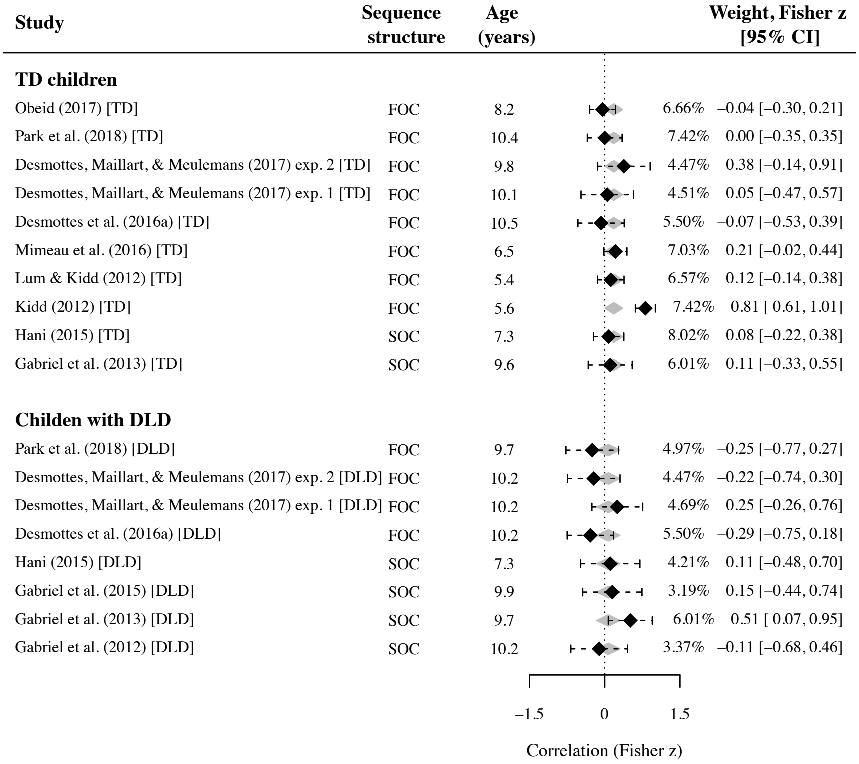


Figure 6.5 Forest plot showing overall and individual mean weighted effect sizes (Fisher *z*) and 95% confidence interval (CI), divided per participant group. The shaded diamonds represent the mean weighted effect size per group (DLD or typically developing). TD = typically developing; FOC = first-order conditional; SOC = second-order conditional.

6.8 Discussion of the Meta-Analysis

The present meta-analysis provided a quantitative overview of published and unpublished studies on the association between serial reaction time performance and expressive grammatical proficiency in children with and without DLD. Summarizing over 18 unique correlations that collectively examined 139 children with DLD (8 effect sizes) and 453 typically developing children (10 effect sizes), we found no evidence for or against the existence of an association between serial reaction time task performance and expressive grammatical proficiency in children with and without DLD. According to the declarative compensation hypothesis, the investigated correlation may be smaller in children with DLD as compared to their typically developing peers, which may result in a weaker overall correlation in the pooled group of children. Therefore, we also assessed whether the mean weighted correlation is smaller in children with DLD than in typically developing children. The latter could not be concluded. In the General discussion we discuss some factors that may have contributed to these inconclusive results.

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

6.9 General discussion

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

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

There are various factors that may have contributed to these inconclusive results. Firstly, they may be (partially) the result of psychometric shortcomings in the currently available measures to assess individual differences in statistical learning (Arnon, 2019; Siegelman, Bogaerts, & Frost, 2017; West et al., 2017). Secondly, studies on the relation between statistical learning and other cognitive processes often spend very little time discussing the theoretical motivation behind the selection of their tasks (as commented on by Siegelman, Bogaerts, Christiansen, & Frost, 2017). As a consequence, the sequential structure targeted in the statistical learning tasks is often only tangentially related to structure relevant for the linguistic ability that researchers try to predict with their task, let alone to how children acquire language in real life. In the set of studies that were included in our meta-analysis, we indeed observe that first-order conditional reaction time tasks and second-order conditional reaction time tasks, which clearly differ in their underlying sequential structure, are used to predict proficiency on the exact same grammar tasks.

Furthermore, most of the grammar tasks assess children's knowledge of a mixture of grammatical structures. The sentence recall task (Semel et al., 2010), for example, measures children's knowledge of different sentence types (e.g. passives, declaratives, relative clause constructions), different morphosyntactic processes (subject–verb agreement, past-tense production, pluralization) and likely also other cognitive processes such as working memory (Frizelle, O'Neill, & Bishop, 2017). However, the pattern that needs to be learned in the serial reaction time task may not be relevant in predicting sensitivity to all these different sentence types and morphosyntactic constructions (Misyak & Christiansen, 2012; Wilson et al., 2018; Mimeau et al., 2016; Kidd & Arciuli, 2016).

Finally, there may also be a discrepancy in how acquired knowledge is measured in statistical learning tasks versus how acquired knowledge is measured in grammatical proficiency tasks. In the present sample of studies, the measures used to assess visuomotoric statistical learning are all processing-based (i.e. based on response times), whereas the measures used to assess grammatical proficiency are all, except for Clark and Lum (2017), accuracy-based. Processing-based measures may be more sensitive to implicit knowledge representations, whereas accuracy-based measures may be more sensitive to explicit knowledge representations (Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015; Misyak et al., 2010; Isbilen, McCauley, Kidd, & Christiansen, 2017). This discrepancy may complicate the detection of an association between the two cognitive systems.

6.10 Conclusion

Neither our own experiment nor our meta-analysis provides any evidence for the existence of an association between serial reaction time performance and expressive grammatical proficiency in children with and without DLD. The confidence interval of the meta-analysis (Pearson *r* from −.038, to +.28) is compatible with a nonexistent association, but also with a medium-sized association. We speculate that such an association may exist only if (a) the targeted structure in the statistical learning task is meaningfully related to the target structure in the grammatical proficiency task *and* (b) both measures represent the same represent the same response type of the participant. Overall, it is even well possible that visuomotoric statistical learning is associated with expressive grammar but that we encountered methodological problems in its detection. Taken together, we cannot claim yet that a visuomotoric statistical learning deficit is or is not associated with the language problems observed in children with DLD.

Acknowledgements

This work is part of the research programme Vernieuwingsimpuls – Vidi (examining the contribution of procedural learning to grammar and literacy in children) with project number 27689005, which is (partly) financed by the Dutch Research Council (NWO), and awarded to Judith Rispens (PI). For the experimental part of this paper, we extend our gratitude to all children who participated and to their parents, teachers, and speech therapists who facilitated the children's participation. More specifically, we would like to thank Viertaal special education; The Royal Dutch Auris Group; The Royal Dutch Kentalis;

Pento, and FOSS/ stichting Hoormij for their help in the recruitment of children with DLD. Also, we would like to thank the four primary schools that participated. We also thank Jarrad Lum and Evan Kidd for sharing their serial reaction time task and Dirk Jan Vet for the technical implementation of the experiment. Finally, we thank Iris Broedelet-Resink, Sascha Couvee, Darlene Keydeniers and Maartje Hoekstra (test assistants) and our student assistant for recording the Dutch version of the serial reaction time task instructions. For the meta-analysis we thank Hailin Hao and Ellen Collée (literature search and coding of variables). Also, we thank Dorothy Bishop, Lise Desmottes, Sengottuvel Kuppuraj, Chirstelle Maillart, Catherine Mimeau, Rita Obeid, Jisook Park, Sybren Spit and Gillian West for providing additional data and information that enabled us to conduct the meta-analysis.

Chapter 7

7.1 General discussion

Children differ in the (apparent) ease with which they acquire language. The studies in this dissertation investigated whether this difference in ease of language acquisition correlates with children's sensitivity to statistical regularities in the input. More specifically, we investigated (1) whether we could detect differences in sensitivity to statistical regularities at the group and individual level (this concerns the measurement of statistical learning), (2) whether individual differences in statistical learning ability correlated with language proficiency and (3) whether the problems observed in children with Developmental Language Disorder (DLD) can be explained by a statistical learning deficit that is observable across modalities (auditory, visual, visuomotor), domains (verbal, nonverbal) and dependency types to be learned (adjacent dependencies, nonadjacent dependencies, mixed adjacent and nonadjacent dependencies). This final chapter provides a summary and synthesis of the individual studies described in this dissertation. The chapter ends with some notes on the clinical relevance of this (type of) research.

7.1.1 Summary of the findings

The main aim of **Chapter 2** was to provide a quantitative overview (meta-analysis) of all auditory verbal statistical learning studies in people with and without DLD. This overview also provided an estimate of the mean weighted difference (effect size) in auditory verbal statistical learning performance between people with and without DLD. This estimate of the DLD–TD difference in auditory verbal statistical learning appeared to be moderate to large and the direction of the difference is compatible with the hypothesis that DLD is associated with an auditory verbal statistical learning deficit: on average people with DLD performed 0.54 standard deviations worse than people without DLD. We could not draw a conclusion about any modulation of the deficit by the linguistic level at which learning took place (word segmentation studies versus grammar learning studies) or the participants' age. An additional benefit of the meta-analysis is that it provided a clear overview of (a) the measurement types that have been used to assess auditory verbal statistical learning and (b) the age groups in which this type of statistical learning has been investigated. Notably, this overview revealed (a) that most studies on auditory verbal statistical learning used only offline measures (e.g., grammaticality judgments) of learning, and (b) that none of the included studies investigated nonadjacent dependency learning in primary school children. It thus remains to be seen if the observed DLD–TD difference can be replicated using online measures of learning and/or using nonadjacent dependencies in primary-school-aged children.

The main aim of **Chapter 3** was therefore to develop an online child-friendly measure of nonadjacent dependency learning that can detect this type of learning in typically developing primary-school-aged children (aged between 5 and 8 years old). Nonadjacent dependency learning is commonly assessed via offline grammaticality judgment measures. The use of such judgments may be problematic in primary-school-aged children as the ability to make a grammaticality judgment likely depends on metalinguistic awareness, an ability that children acquire relatively late. Also, offline measures reflect only the outcome of the learning process, disregarding information on children's learning trajectory. Using our novel online measure, we showed a difference in children's response times to structured items that were predictable (due to their being part of a nonadjacent dependency relation) and their response times to unstructured items that were unpredictable (henceforth this difference in response times is referred to as "disruption peak"). The results of our offline measure of nonadjacent dependency learning (grammaticality judgment) did not provide evidence for or against children's learning of the dependencies. In the next section of this discussion (Measuring statistical learning) we discuss the use of online and offline measures of statistical learning.

In **Chapter 4** we investigated whether the size of the disruption peak (see Chapter 3) is smaller in children with DLD as compared to their typically developing peers. If so, this would be in agreement with the hypothesis that children with DLD between 8 and 12 years have an auditory verbal nonadjacent dependency learning deficit. We did observe that children with DLD had smaller disruption peaks than their typically developing peers. We could not detect children's learning of the dependencies with our offline measure of learning. Also, we observed no evidence for or against an association between our online measure of nonadjacent dependency learning and grammatical proficiency, as measured with the sentence recall task and word structure task of the Clinical Evaluation of Language Fundamentals – Dutch version (CELF, Semel, Wiig, & Secord, 2010).

We did not assess the correlation between our offline measure of statistical learning and grammatical proficiency, as we did not detect learning with this offline measure.

The main objective of **Chapter 5** was to assess whether children with DLD have a visual nonverbal statistical learning deficit. Assuming that children with DLD have a statistical learning deficit that is independent of the modality and domain in which learning takes place, we hypothesized to observe such deficit. Furthermore, as in typically developing children it has been shown that visual statistical learning correlates with literacy performance, we were also interested to see whether individual differences in visual nonverbal statistical learning between children with DLD correlated with individual differences in literacy proficiency. Such an association may in part explain the individual differences in literacy performance among our children with DLD: approximately half of the children with DLD had difficulties reading and spelling. The visual nonverbal statistical learning task was set up such that differences in transitional probabilities indicated which three aliens formed a triplet (and thus always came together). If a child is sensitive to these differences in transitional probabilities, then s/he learns the triplet structure. Using offline measures of learning (triplet completion task and triplet recognition task) we found no evidence for or against a difference in learning between children with and without DLD. Interestingly, the offline measures provided evidence that children with DLD were sensitive to the triplet structure to be learned. Using online measures of learning, no learning effect was detected in children with and without DLD. Finally, we found no evidence for (against) an association between visual nonverbal learning and literacy performance (one-minute word reading, Brus & Voeten, 1969; two-minute pseudoword reading, van den Bos, Spelberg, Scheepstra, & de Vries, 1994; spelling, Braams & de Vos, 2015) in children with DLD.

In **Chapter 6** we combined a conceptual replication (experimental study) with a meta-analysis to evaluate the evidence for the proposed association between children's detection of sequential regularities in the nonverbal visuomotoric domain (serial reaction time task) and grammatical proficiency. With the replication study, we found no evidence for (or against) the existence of a visuomotoric statistical learning deficit in children with DLD. Using a meta-analytic approach to further investigate the link between children's performance on a serial reaction time task and grammatical proficiency, we found no evidence for (or against) the existence of such association.

7.1.2 Measuring statistical learning (aim 1)

A recurring issue throughout the individual chapters of this dissertation is the sensitivity of the currently available measures of statistical learning at the group level and at the individual level. This section aims to synthesize the points raised concerning this topic as well as to evaluate our use of novel child-friendly measures (Chapters 3, 4 and 5) of statistical learning. As already discussed in Chapters 3 and 4, we conclude that the use of online measures of statistical learning in addition to the use of offline measures is an advancement, because the measurement types may tap into different kinds of knowledge representations. Batterink and Paller (2019), for example propose that statistical learning performance comprises at least two dissociable components: (1) perceptual binding and (2) subsequent memory storage and retrieval. Perceptual binding happens online while participants are exposed to the stimuli and can thus be best measured with online measures of learning. Offline measures of learning may be more sensitive to the second component (memory storage and retrieval) of statistical learning. Relatedly, it has also been proposed that online measures of statistical learning can be best described as processing-based measures of learning whereas the offline measures can be best described as reflection-based measures of learning (Isbilen, Frost, Monaghan, & Christiansen, 2018).

At the group level, we detected statistical learning using an online measure of auditory verbal statistical learning (Chapters 3 and 4), offline measures of visual nonverbal statistical learning (Chapter 5) and an online measure of visuomotor nonverbal statistical learning (Chapter 6). Unique in this dissertation is that we have online measures for each of these three types of statistical learning. While the use of online measures is standard in serial reaction time studies (Chapter 6), the use of such measures is novel, especially with primary-school-aged children, in auditory nonadjacent dependency learning studies (Chapters 3 and 4) and visual statistical learning studies (Chapter 5, and see also van Witteloostuijn, Lammertink, Boersma, Wijnen and Rispens, 2019).

As the use of online measures of statistical learning is relatively new, novel methods keep emerging. This is also why we used two different online measurement types in this dissertation. In Chapter 5, children’s difference in response times between predictable and less predictable elements (“RT predictability advantage”) was taken as a measure of learning in the visual nonverbal task. The difference in children’s response times to unstructured trials as compared to structured trials ("RT disruption peak") functioned as a measure of learning in the auditory verbal task (Chapters 3 and 4) and the visuomotor nonverbal task (Chapter 5). Interestingly, we detected a learning effect in the tasks that used the disruption peak measure, but could not detect learning using the RT predictability advantage measure. From this observation, we speculate that the RT disruption measure may be more sensitive in its detection of statistical learning than the RT predictability advantage measure. This may be because the decrease in response time – as assessed with the RT predictability advantage measure – does not necessarily reflect statistical learning only: children may also become faster as a result of task adaptation or increasing familiarity with the stimuli (Karuza, Farmer, Fine, Smith, & Jaeger, 2014; Kidd & Kirjavainen, 2011; but see Kuppuraj, Duta, Thompson, & Bishop, 2018 for a potential solution to this problem). With this type of measure, it is therefore difficult to disentangle statistical learning from general effects of practice. However, it should be noted that we cannot conclude that this apparent difference in sensitivity between both measurement types is real, because we did (and could) not directly compare the outcomes of the three statistical learning tasks in one statistical model.

Other than a difference in measurement sensitivity, the apparent difference may also be a side effect of differences in task design between the visual statistical learning task (RT predictability advantage measure) and the two tasks that used the RT disruption measure (auditory verbal statistical learning task; serial reaction time task). In all three tasks, we instructed children to respond as quickly as possible. However, the number of response options differed per task. In the visual task there was only one possible answer (the spacebar), the auditory task had two response options (green button or red button) and the visuomotor task had four response options (four locations on the screen). This also means that children's responses in the visual task do not reflect accuracy, as there was no 'target' answer. This is different from the auditory task and visuomotor task where children's responses reflect accuracy. This difference makes that an incorrect answer has consequences in the latter two tasks, but not in the visual task. A consequence of this difference is that a predictive strategy, that is predicting or anticipating which button to press, may be more beneficial for the auditory and visuomotoric task than for the visual task. An incorrect answer has consequences in the first two tasks, but not in the visual task. This may mean that children may not have used a predictive strategy in the visual statistical learning task, which makes it then difficult to detect a corresponding "predictability advantage".

Complementary to the online measures of statistical learning, our auditory nonadjacent dependency learning task (Chapters 3, 4) and our visual statistical learning task (Chapter 5) also assessed children's sensitivity to structure with offline measures of learning. We did detect a learning effect using the offline measures of visual statistical learning. We could not detect a learning effect using the offline measures of auditory nonadjacent dependency learning. Again, as we did not compare sensitivity to the visual and auditory structures using offline measures in one statistical model, future work is needed to determine whether the measures indeed differ in their sensitivity to detect statistical learning. It may be remarkable in this context, however, that the offline measures of our visual statistical learning task followed recommendations given by Siegelman, Bogaerts and Frost (2017) to enhance their sensitivity. In these measures, we increased the number of test items and used different types of offline measures (triplet completion and triplet recognition; Chapter 5). We did, however, not implement these recommendations in the offline measure of our auditory nonadjacent dependency learning tasks (Chapters 3 and 4), which may have made the offline measures used in these chapters less sensitive than the offline measures used in the visual statistical learning task (Chapter 5).

Another explanation for the apparent (but not confirmed) difference in learning effects detected in the two offline measures has to do with the type of acquired knowledge that offline measures of statistical learning are assumed to be sensitive to. It has been proposed that offline measures of statistical learning appeal to explicit or metalinguistic knowledge of the structure (Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015). As discussed in Chapter 5, we have reasons to believe that, in comparison to other work on visual statistical learning in children with DLD (Noonan, 2018) and compared to our own auditory nonadjacent dependency learning task (Chapters 3 and 4), the set-up of our visual statistical learning task may have triggered a more explicit learning strategy. This is mostly because, in comparison to the other studies, the task instructions of our visual statistical learning task were explicit in telling the children that they should pay attention to the order in which the aliens appeared. Recently, Himberger, Finn and Honey (preprint) showed that, using offline measures of learning, adults' performance on a visual statistical learning task improved when the adults received the explicit instruction to search for the regularities as compared to when they did not receive such explicit instruction. These results indicate that the explicitness of the instruction may impact learning and should be considered when designing a statistical learning task. Again, we would like to stress that our reasoning is speculative and far from conclusive, particularly also because the effect of such explicit instructions to focus on structure in this type of tasks is likely to be different for adults than for children (if nothing else because children may not really understand the concept “structure” or “order” yet).

In addition to investigating statistical learning at the group level, we also aimed to detect statistical learning at the individual level. To this end, we extracted individual measures of the online and offline measures that were described above (see Chapters 4, 5 and 6 for procedures). We needed these individual measures to assess the strength of the association between statistical learning and language proficiency, as discussed later on (see next section, aim 2). However, an association between individual measures of statistical learning and language proficiency can only be detected with high between-subject variability in both measures. This is in conflict with our aim to detect group level differences, because in order to detect differences at the group level, the between subject variability should be low (Hedge, Powell, & Summer, 2017 and see Siegelman, Bogaerts, & Frost, 2017; West, Vadillo, Shanks, & Hulme, 2017 for similar conclusions within the statistical learning literature). This may be one of the reasons why we could not detect an association between statistical learning and linguistic proficiency in any of our individual studies (Chapters 4,5,6). That is, our individual measures of statistical learning may have been psychometrically weak in their assessment of learning.

Alternatively, it may also be that we could not detect an association between our online measures of statistical learning and language proficiency, because of our assumptions on how individual sensitivity to regularities is expressed were incorrect. This is analogous to a point of contention in the infant literature where it is frequently doubted at what scale infants’ looking time or listening time preferences represent individual differences (Durrant, Jessop, Chang, Bidgood, Peter, Pine, & Rowland, preprint). Using the size of children’s disruption peak measure (or looking/listening time preference in the infant literature) we assume a linear relationship between the size of the disruption peak and the child’s sensitivity to the regularities. That is, a child with a disruption peak of 40 milliseconds is considered as being twice as sensitive to the regularities than a child with a disruption peak of 20 milliseconds. It is questionable, however, if such numerical difference is meaningful at all. It could also be the case that a categorical distinction is more in place. Such categorical distinction requires a certain threshold, for example if children have disruption peaks larger than 20 milliseconds then they are sensitive to the regularities whereas children who have peaks smaller than 20 milliseconds may have not learned the regularities. It could be interesting for future work to investigate whether such a threshold can be determined and to simulate whether a child's sensitivity to statistical regularities is best expressed in a linear or categorical way (for such a simulation on infant's dynamic event understanding see Durrant et al., preprint).

7.1.3 Statistical learning and language proficiency (aim 2)

The second aim of this dissertation was to assess the association between statistical learning, grammar and literacy proficiency in children with and without DLD. In Chapters 4 and 5, this assessment was secondary to the assessment of the deficit itself (and thus exploratory). In Chapter 6 the assessment of this association was part of our confirmatory analysis. Unfortunately, the outcomes of both exploratory analyses and the confirmatory analysis are inconclusive. This may mean that the association may be weak or nonexistent (Chapters 4, 5 and 6), but also that the association does exist and may even be strong (Chapter 6). In Chapter 6 and the section above, we already introduced several methodological issues that may have hampered the detection of the hypothesized association. We concluded (1) that the measures of statistical learning may be psychometrically weak in their assessment of individual differences as they are designed to detect differences at the group level and (2) that the assumption of a linear relationship between children's size of the disruption peak (or scores at the offline test) and their sensitivity to statistical structures may be invalid. Besides these methodological issues, we also speculated (Chapter 6) that the a-specific nature of our language proficiency measures may have hampered the detection of the proposed association. That is, the targeted structure in the statistical learning task may not have been related to the (various) targeted structure(s) in the language proficiency tasks (also commented on in Siegelman, Bogaerts, Christiansen, & Frost, 2017). Following the predictions of general statistical learning accounts, one would expect associations between statistical learning and measures of language proficiency to be independent from the structures used. Therefore, if it is true that associations between statistical learning and language only exist when the targeted structures in both tasks are similar then this poses a problem for general accounts of statistical learning.

In addition to the outcomes of our individual studies on the association between statistical learning and language proficiency, we also reported the outcomes of a meta-analysis that analysed the outcomes of different studies on the serial reaction time – grammatical proficiency association (Chapter 6). One advantage of such a meta-analytic approach is that it allows for the identification of moderators of the effect that may be difficult to assess with one single study. In our particular case, the individual studies that were included in the meta-analysis covered a range of different ages (children between seven and twelve years of age), the use of different sequence structures (first-order conditional and second-order conditional) as well as two different populations (children with and without DLD). This allowed us to explore whether any of these moderators influenced the strength of the serial reaction time – grammatical proficiency association. The results of the analyses were inconclusive and thus we could not draw a conclusion about any modulation of the association by participants' age, sequence type or population.

7.1.4 A statistical learning deficit in children with DLD (aim 3)

The experimental studies described in Chapters 4, 5 and 6 assessed the existence and size of a statistical learning deficit in primary-school-aged children with DLD across three different paradigms. As summarized at the start of this chapter, we observed a DLD–TD difference on the auditory verbal nonadjacent dependency learning task (Chapter 4). This difference led to the conclusion that children with DLD have an auditory verbal statistical learning deficit. We could not conclude that children with DLD have a statistical learning deficit outside the auditory verbal domain: the results of Chapters 5 and 6 provided no evidence for or against a visual nonverbal statistical learning deficit (Chapter 5) or a visuomotor nonverbal statistical learning deficit (Chapter 6). In these latter two chapters it was observed that, when using an offline measure of learning (Chapter 5) and when using an online measure of learning (Chapter 6), children with DLD were sensitive to the to-be-learned structures they had been exposed to. Although it is appealing to conclude that this pattern of results shows that the deficit is restricted to the auditory verbal domain, such a conclusion is premature, as we cannot directly compare the outcomes of the three individual studies. All three statistical learning tasks targeted a different sequential structure, that is the auditory task targeted nonadjacent dependencies, the visual task targeted adjacent dependencies, and the visuomotoric task targeted a fixed sequence comprising both adjacent and nonadjacent dependencies. This use of different sequential structures impedes a comparison across the studies, as recent work suggests that the specific structure to be learned may impacts its learnability. For example, the detection of nonadjacent dependencies is thought to be more cognitively demanding than the detection of adjacent dependencies (Wilson et al., 2018) and may also result in more explicit knowledge representations (Romberg and Saffran, 2013). This means that if one wants to draw any conclusions on the domain-, or modality-specific constraints of the statistical learning deficit, one should keep the targeted structure constant. For example, future studies could compare our auditory nonadjacent dependency learning task (Chapters 3 and 4) to a visual linguistic nonadjacent dependency learning task as described in Karuza et al. (2014).

Another reason that makes it difficult to disentangle modality specific effects from domain specific effects on the presence and size of a statistical learning deficit in children with DLD is that only our auditory task included linguistic stimuli, while both non-auditory tasks used nonverbal stimuli. In hindsight we realize that it is therefore difficult to rule out the possibility that rather than an auditory verbal statistical learning deficit, the DLD–TD difference observed in Chapter 4 is the result of an auditory processing difficulty (Tallal, 2000) or reduced linguistic entrenchment (Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018) in children with DLD as compared to their typically developing peers. This issue may be resolved by future studies in which statistical learning in children with and without DLD is compared across four different tasks: (a) an auditory verbal statistical learning task, (b) an auditory nonverbal statistical learning task, (c) a non-auditory verbal statistical learning task and (d) a non-auditory nonverbal statistical learning task (see Evans, Saffran and Robbe-Torres, 2009 or Noonan, 2018 for other combinations of statistical learning task comparisons in people with and without DLD).

Part of the explanatory power of the statistical learning deficit hypothesis lies in its assumed multi-component nature. It has been hypothesized that successful statistical learning depends on other cognitive capacities such as processing speed, various forms of attention and different types of memory (Arciuli, 2018). There is some evidence, for example, that learning rules from speech is a two-stage process: after statistical regularities have been detected, learning shifts towards a more goal-directed attentional stage. In this second stage, learners integrate *what* is learned and *when* is learned in a goal-directed manner, which in turn facilitates generalization of the rules (Orpella et al., 2019). Most children with DLD have reduced processing speed, working memory skills and attention skills as compared to their typically developing peers. Therefore, the statistical learning deficit hypothesis fits well with the heterogeneous profile of problems observed in these children. In a laboratory setting, however, one may aim for a pure measure of children's sensitivity to regularities and therefore decide to control for any group differences in the cognitive areas that may support statistical learning. This was also the reason that we controlled for verbal working and verbal short-term memory in Chapter 4. In hindsight one may wonder whether controlling for these memory types hampers the ecological validity and generalizability of the outcomes and it could potentially explain why the size of our auditory verbal statistical learning deficit (Chapter 4) is relatively small in comparison to other statistical learning studies in the auditory verbal domain (Chapter 2). Most studies included in the meta-analysis (Chapter 2) did not control for potential other cognitive processes that may have had an impact on the size of the statistical learning deficit.

One of the advantages of our meta-analysis (Chapter 2) on the auditory verbal statistical learning deficit in DLD is that we obtained an estimate of the size of the auditory verbal statistical learning deficit in people with DLD and that we could thus interpret the outcomes of our individual studies in light of this effect size. The estimate of our point estimate of standardized effect size for the between group difference on our online measure of auditory verbal statistical learning is 0.23. The point estimate of this difference on our visuomotor statistical learning task is 0.022. Both these estimates are smaller than the lower bound of the confidence interval for the mean weighted effect size, which is 0.36 (across 10 studies) reported in Chapter 2. This means that our observed DLD–TD differences are relatively small. Interestingly, all studies, except one, that were included in the meta-analysis reported in Chapter 2, used only offline measures of statistical learning. As the use of online measures of statistical learning is becoming more common it would be interesting to have more studies with online measures of a DLD–TD difference. These studies could then be added to our meta-analysis (as our meta-analysis is community-augmented and thus open to updates, see Chapter 2 for details) and in the future we can then assess whether the size of the deficit is modulated by measurement type (offline versus online measure). Note that we did not discuss the DLD–TD differences of our offline measures of visual nonverbal statistical learning. This is because we interpreted these differences in terms of odds ratio effect sizes and as yet, there is no general consensus on how to interpret the magnitude of odds ratio effect sizes (but see Chen, Cohen and Chen, 2010). Also, it may be good to note that we do not interpret the effect sizes of our DLD–TD difference for our offline measures of auditory verbal statistical learning (Chapter 4) and our online measures of visual nonverbal statistical learning (Chapter 5). This is because we did not detect learning with these measures.

7.1.5 Clinical implications

In the long run, research into the role of more general cognitive processes, such as statistical learning, that are thought to be associated with language proficiency, may have value in the context of diagnosing and treating DLD. It has been suggested that training of such cognitive processes contributes to the success of language treatment programs in children with developmental language problems (Montgomery, Magimairaj, & Finney, 2010; Plante & Gómez, 2018).

The small magnitude of our observed auditory verbal statistical learning deficit (Chapter 4) together with the weak associations between statistical learning, grammatical proficiency (Chapter 4) and literacy (Chapter 5) suggest that interventions aimed at bolstering children's statistical learning will have limited, if any, effects. As already concluded in Chapter 4, it may be more effective to focus on the training of other more general cognitive processes, such as phonological processing or working memory, that are potentially more strongly correlated to children's language proficiency than statistical learning. For example, a meta-analysis on differences in nonword repetition between children with and without DLD reported that children with DLD performed on average 1.27 standard deviations below children without DLD (Graf Estes, Evans and Else-Quest, 2007).

This is not to say, however, that interventions for DLD should disregard the principles of statistical learning. Plante and Gómez (2018) explain that it is relatively easy to incorporate statistical learning principles into existing treatments for DLD. Incorporation of these principles may lead to enhanced learning of morphological target structures in children with DLD (Plante et al., 2014). The concrete example that Plante and Gómez (2018) describe is that if treatment aims at children's correct use of the grammatical third person -s morpheme, then detection of the dependency between the subject and the -s morpheme is facilitated if the combination is used in different contexts, that is with many different verbs. In their intervention study, Plante et al. (2014) showed that children with DLD who received treatment of a grammatical morpheme (e.g., third person -s) in a high variability context (24 unique verbs) produced more correct forms of the trained grammatical morpheme than children in a low variability context (12 unique verbs).

In the context of more metalinguistic learning strategies, the detection of structure in children with DLD may benefit from the inclusion of visual cues that explicitly draw children's attention to the underlying linguistic structure (e.g., Ebbels, 2007). Our observation that children with DLD were sensitive to structure in the visual task may indirectly support the use of such strategy. Note that we addressed none of the discussed clinical implications in the present dissertation, however. Therefore, more research is needed to confirm that indeed the acquisition of grammatical morphemes may benefit from the proposed strategies that aim to facilitate children's detection of structure.

7.2 Conclusion

The discussion above made clear that the size of the statistical learning deficit in children with DLD as well as the strength of the association between statistical learning and language proficiency may depend on several factors, including but not restricted to the domain and modality in which learning takes place, the specific structure to be learned and the way in which statistical learning is measured. The quantitative overview on auditory verbal statistical learning (Chapter 2) and our experimental study on auditory verbal nonadjacent dependency learning (Chapter 4) show that people with DLD are less sensitive to regularities in the auditory verbal domain than people without DLD. We could not conclude that children with DLD have (or do not have) a statistical learning deficit outside this domain (Chapters 5 and 6) nor that statistical learning ability correlates or does not correlate with grammar and literacy proficiency (Chapters 4, 5 and 6). Although tempting, it is premature to conclude from the present set of results that people with DLD have a statistical learning deficit that restricts itself to the auditory verbal domain. More research is needed to confirm that the observed difference between people with and without DLD in this domain is indeed the consequence of reduced statistical learning and not of deficiencies in other cognitive areas such as auditory processing or reduced linguistic entrenchment in people with DLD. This question may be answered if future studies adapt our novel online measure, that has been shown to reliably detect children's auditory verbal nonadjacent dependency learning (Chapters 3 and 4), to comparable tasks in the auditory nonverbal, visual linguistic and visual nonverbal domain.

Epilogue

The first time I read about the "word counting watch" in Dave Eggers' novel *The Circle* (see Prologue), I was sceptical: part of the beauty of children's first language acquisition process is that it happens in such a natural and uncontrolled way. After four years of study, I still believe that it should never be a goal to fully control, regulate and monitor this process. Nevertheless, a tool like Eggers' watch may open new avenues in our possibilities to investigate the interaction between children's natural language input and their cognitive capacity to process this input. Therefore, one of the things that I would really like to further investigate is whether children with DLD indeed benefit from language input that is delivered in a more controlled and structured way. In such context Eggers' watch may have the potential to become a valuable tool for researchers and professionals working with children with DLD.

References

References marked with an asterisk (*) indicate studies included in the meta-analysis on auditory verbal statistical learning (Chapter 2).

References marked with double asterisks (**) indicate studies included in the meta-analysis on the statistical learning-expressive grammar correlation (Chapter 6).

References marked with triple asterisks (***) indicate studies included in the meta-analysis on statistical learning-receptive grammar correlation (Reported on OSF only, see Chapter 6).

References marked with quadruple asterisks (****) indicate studies included in both the meta-analysis on the statistical learning-expressive grammar correlation and the meta-analysis on the statistical learning-receptive grammar correlation (Chapter 6).

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

Archibald, L., & Gathercole, S. (2006). Short-term and working memory in specific language impairment. *International Journal on Language & Communication Disorders, 41*(6), 675–693. doi:10.1080/13682820500442602

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

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

Arciuli, J., & Conway, C. (2018). The promise – and challenge – of statistical learning for elucidating atypical language development. *Current Directions in Psychological Science, 27*(6), 1–9. doi:10.1177/0963721418779977

Arciuli, J., & Simpson, I. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Psychology, 14*(3), 464–473. doi:10.1111/j.1467–7687.2009.00937.x

Arciuli, J., & Simpson, I. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286–304. doi:10.1111/j.1551-6709.2011.01200.x

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

Aslin, D., & Newport, E. (2014). Distributional language learning: Mechanisms and models for category formation. *Language Learning, 64*, 85–105. doi:10.1111/lang.12074

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for 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. doi:10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. (2018, May 26). Parsimonious mixed models. Retrieved from the arXiv database: 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), 1–48. doi:10.18637/jss.v067.i01

Batterink, L., & Paller, K. (2019). Statistical learning of speech regularities can occur outside the focus of attention. *Cortex, 115,* 56–71. doi:10.1016/j.cortex.2019.01.013

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

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

Bishop, D. (2014). Ten questions about terminology for children with unexplained language problems. *International Journal of Language and Communication Disorders, 49*(4), 381–415. doi:10.1111/1460-6984.12101

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. doi:10.1111/jcpp.12721

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

Bogaerts, L., Franco, A., Favre, B., & Rey, A. (2016, June). *Speech onset latencies as an online measure of regularity extraction.* Poster session presented at the Fifth Implicit Learning Seminar, Lancaster, England.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis.* Chichester, England: John Wiley & Sons.

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

Brown, L., Sherbenou, R., & Johnsen, S. (1997). Test of Nonverbal Intelligence–Third Edition [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

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

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics Simulation and Computation, 39*(4), 860–864. doi:10.1080/03610911003650383

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.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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

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

Clashen, H., & Hansen, D. (1997). The grammatical agreement deficit account in specific language impairment: Evidence from therapy experiments. In M. Gopnik (Ed.), *The inheritance and innateness of grammar* (pp. 141–160). Oxford, England: Oxford University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*–Second Edition. Hillsdale, NJ: Lawrence Erlbaum Associations.

Cohen, J. (1992). A power primer. *Psycholinguistic Bulletin, 112*(1), 155–159.

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

***Conti-Ramsden, G., Ullman, M., & Lum, J. (2015). The relation between receptive grammar and procedural declarative, and working memory in specific language impairment. *Frontiers in Psychology, 6*:1090, 1–11. doi: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 [Invited commentary on "The procedural learning deficit hypothesis of language learning disorders: we see some problems" by G. West, M. Vadillo, D. Shanks, & C. Hulme]. *Developmental Science, 22*(4):e12814. doi:10.1111/desc.12814

Conway, C., & Christiansen, M. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science, 17*, 905–912. doi:10.1111/j.1467-9280.2006.01801.x

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy, 21*, 648–667. doi:10.1111/infa.12127

Csányi, I. (1974). Peabody Szókincs-Teszt [Peabody Vocabulary Test]. Budapest, Hungary: Bárczi Gusztáv Gyógypedagógiai Foiskola.

Del Re, A. (2013). Compute.es: Compute effect sizes. [R package, version 0.2-2]. Retrieved from http://cran.r-project.org/web/packages/compute.es

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

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

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

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

Dickersin, K. (2005). Recognizing the problem, understanding its origins and scope, and preventing harm. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11–34). Chichester, England: Wiley & Sons.

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

Dunn, M., & Dunn, L. (1981). Peabody Picture Vocabulary Test–Revised [Measurement Instrument]. Circle Pines, MN: American Guidance Service.

Dunn, M., & Dunn, L. (2007). Peabody Picture Vocabulary Test–Fourth Edition. [Measurement Instrument] Circle Pines, MN: American Guidance Service.

Durrant, S., Jessop, A., Chang, F., Bidgood, A., Peter, M., Pine, J., & Rowland, C. (2019). Does the understanding of complex dynamic events at 10 months predict vocabulary development? Manuscript submitted for publication. Retrieved from https://osf.io/mjv73/

Ebbels, S. (2007). Teaching grammar to school-aged children with specific language impairment using shape coding. *Child Language Teaching and Therapy, 23*(1), 67–93. Distributed training enhances implicit sequence acquisition in children with specific language impairment. doi:10.1191%2F0265659007072143

Ebert, K., & Kohnert, K. (2011). Sustained attention in children with primary language impairment: A meta-analysis. *Journal of Speech, Language and Hearing Research, 54*(5), 1372–1384. doi:10.1044/10924388(2011/10-0231)

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

Eggers, D. (2013). *The Circle: A novel.* London, England: Penguin Books 2014.

Eimas, P., Siqueland, E., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 181*, 303–306. doi:10.1126/science.171.3968.303

Elleman, A., Steacy, L., & Compton, D. (2019). The role of statistical learning in word reading and spelling development: More questions than answers. *Scientific Studies of Reading, 23*(1), 1–7. doi:10.1080/10888438.2018.1549045

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

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

*Evans, J., Hughes, C., Hughes, D., Jackson, K., & Fink, T. (2010, June). *SLI—A domain specific or domain general implicit learning deficit? Modality-constrained statistical learning of auditory and perceptual motor sequences in SLI.* Poster session presented at the symposium on research in child language disorders, Madison, WI.

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

Fiser, J., & Aslin, R. (2002). Statistical learning of new feature combinations by infants. *Proceedings of the National Academy of Sciences of the Unites States of America,* 99, 15822–15826. doi:10.1073/pnas.232472899

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:10.1027/1618-3169/a000295

Frank, M., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*, 421–435. doi:10.1111/infa.12182

Friederici, A., Mueller, J., & Oberecker, R. (2011). Precursors to natural grammar learning: Preliminary evidence from 4-month-old infants. *PLoS ONE, 6*(3):e17920. doi:10.1371/journal.pone.0017920

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. doi:10.1017/S0305000916000635

Frost, R., Armstrong, B., & Christiansen, M. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin.* Advance online publication. doi:10.1037/bul0000210

Frost, R., Armstrong, B., Siegelman, N., & Christiansen, M. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19*(3), 117–125. doi:10.1016/j.tics.2014.12.010

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. doi: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. doi:10.1017/S0142716413000490

****Gabriel, A., Maillart, C., Stefaniek, N., Lejeune, C., Desmottes, L., & Meulemans, T. (2013). Procedural learning in specific language impairment: Effects of sequence complexity. *Journal of the International*

Neuropsychological Society, 19, 164–271. doi:10.1017/S1355617712001270

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

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76. doi:10.1016/S0010-0277(98)00034-1

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

Gómez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135. doi:10.1016/S0010-0277(99)00003-7

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*, 183–206. doi:10.1207/s15327078in0702_4

Graf Esters, K., Evans, J., Else-Quest, N. (2007). Differences in nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 50*(1), 177–195. doi:10.1044/1092-4388(2007/015)

Grama, I., Kerkhoff, A., & Wijnen, F. (2016). Gleaning structure from sound: The role of prosodic contrast in learning non-adjacent dependencies. *Journal of Psycholinguistic Research, 45(*6), 1427–1449. doi:10.1007/s10936-016-9412-8

*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. doi: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. doi:10.1111/jcpp.12734

Hamrick, P., Lum, J., & Ullman, M. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences of the Unites States of America, 115*(7), 1487–1492. doi:10.1073/pnas.1713975115

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

Hedenius, M., Persson, J., Alm, P., Ullman, M., Howard Jr, J., Howard, D., & Jennische, M. (2013). Impaired implicit sequence learning in children with developmental dyslexia. *Research in Developmental Disabilities, 34*(11), 3924–3935. doi:10.1016/j.ridd.2013.08.014

Hedge, C., Powell, G., & Summer, P., (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavioral Research Methods, 50*(3), 1166–1186, doi:10.3758%2Fs13428-017-0935-1

Henderson, L., & Warmington, M. (2017). A sequence learning impairment in dyslexia? It depends on the task. *Research in Developmental Disabilities, 60*, 198–210. doi:10.1016/j.ridd.2016.11.002

Higgins, J., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions*, version 5.1.0 (updated March 2011). Retrieved from http://www.handbook.cochrane.org

Hill, E. (2001). Non-specific nature of specific language impairment: A review of the literature with regard to concomitant motor impairments. *International Journal of Language and Communication Disorders, 36*(2), 149–171. doi:10.1080/13682820118418

Himberger, K., Finn, A., & Honey, C. (2019). Reconsidering the automaticity of visual statistical learning. Manuscript submitted for publication. Retrieved from the arXiv database: https://psyarxiv.com/r659w

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

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

Hsu, H., & Bishop, D. (2014b). Training understanding of reversible sentences: A study comparing language-impaired children with age-matched and grammar-matched controls. *PeerJ*, *3*, 1–23. doi: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*:175, 1–10. doi: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, Language and Hearing Research, 60*(3), 694–700. doi:10.1044/2016_JSLHR-L-150232

Isbilen, E., Frost, R., Monaghan, P. & Christiansen, M. (2018). Bridging artificial and natural language learning: Comparing processing- and reflection-based measures of learning. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1856–1861). Austin, TX: Cognitive Science Society.

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, England.

Joye, N., Broc, L., Olive, T., & Dockrell, J. (2019). Spelling performance in children with developmental language disorder: A meta-analysis across European languages. *Scientific Studies of Reading*, *23*(2), 1–32. doi:10.1080/10888438.2018.1491584

Karuza, E., Farmer, T., Fine, A., Smith, F., & Jaeger, T. (2014, July). On-line measures of predication in a self-paced statistical learning task. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, Canada, 36*, 725–730, Retrieved from https://escholarship.org/uc/item/9s07x343

Kas, B., & Lukács, Á. (2011). Magyar Mondatutánmondási Teszt [HungarianSentence Repetition Test].

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

Kerkhoff, A., de Bree, E., & Wijnen, F. (submitted). Dyslexia, language, and statistical learning. In T. Mintz (Ed.), *Current trends in statistical approaches to language acquisition.* Abingdon, England: Taylor & Francis.

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

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

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

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. doi:10.1080/01690965.2010.493735

Kirkham, N., Slemmer, J., & Johnson, S. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), 35–42. doi:10.1016/S00100277(02)00004-5

Krogh, L., Vlach, H., & Johnson, S. (2013). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology, 3*:598. doi:10.3389/fpsyg.2012.00598

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*(4), 1326–1340. doi: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*(2):171678. doi:10.1098/rsos.171678

***Kuppuraj, S., Rao, P., & Bishop, D. (2016). Declarative capacity does not trade-off with procedural capacity in children with specific language impairment. *Autism & Developmental Language Impairments, 1*, 1–17. doi:10.1177/2396941516674416

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017 [Chapter 2 of this dissertation]). Statistical learning in specific language impairment: A meta-analysis. *Journal of Speech, Language and Hearing Research, 60*(12), 3474–3486. doi:10.1044/2017_JSLHR-L-16-0439

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019 [Chapter 4 of this dissertation]). Children with developmental language disorder have an auditory verbal statistical learning deficit: Evidence from an online measure. *Language Learning, 70*(1), 137–178. doi:10.1111/lang.12373

Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020 [Chapter 5 of this dissertation]). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing: An Interdisciplinary Journal.* Advance online publication: doi:10.1007/s11145-020-10018-4

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (under review [Chapter 6 of this dissertation]). Statistical learning in the visuomotor domain and its relation to grammatical proficiency in children with and without DLD: A conceptual replication and meta-analysis.

Lammertink, I., van Witteloostuijn, M., Boersma, P., Wijnen, F., and Rispens, J. (2019 [Chapter 3 of this dissertation]). Auditory statistical learning in children: Evidence from an online measure. *Applied Psycholinguistics, 40*(2), 279–302. doi:10.1017/S0142716418000577

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.

Lenneberg, E. (1967). The biological foundations of language. *Hospital Practice, 2*(12), 59–67.

Leonard, L. (2014). *Children with specific language impairment*. Cambridge, MA: MIT Press: ISBN: 978-0-262-02706-9.

Leslie, L., & Caldwell, J. (1995). *Qualitative reading inventory*–Second Edition. New York, NY: Addison-Wesley.

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal grammar and statistical learning. *The Annual Review of Linguistics, 1,* 333–353. doi:10.1146/annurev-linguist-030514-125236

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

Lukács, Á., Győri, M., & Rózsa, S. (2012). A TROG pszichometriai jellemzoinek magyar vizsgálata, a normák kialakítása [The psychometric analysis of

Hungarian data from the TROG]. In D. Bishop (Ed.), *TROG—Test for reception of grammar handbook* (pp. 47–86). Budapest, Hungary: OS Hungary.

*Lukács, Á., & Kemény, F. (2014). Domain-general sequence learning deficit in specific language impairment. *Neuropsychology, 28*(3), 472–483. doi:10.1037/neu0000052

Lukács, Á., & Kemény, F. (2015). Development of different forms of skill learning throughout the lifespan. *Cognitive Science, 39*, 383–404. doi:10.1111/cogs.12143

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

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. doi:10.1016/j.cortex.2013.10.011

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

Lum, J., Lammertink, I., Clark, G., Fuelsher, I., Hyde, C., Enticott, P., & Ullman, M. (2019). Visualspatial sequence learning on the serial reaction time task modulates the P1 event-related potential. *Psychophysiology, 56*(2), 1–12. doi:10.1111/psyp.13292

Lum, J., Ullman, M., & Conti-Ramsden, G. (2013). Procedural learning is impaired in dyslexia: Evidence from a meta-analysis of serial reaction time studies. *Research in Developmental Disabilities, 34*, 3460–3476. doi:10.1016/j.ridd.2013.07.017

Mainela-Arnold, E., & Evans, J. (2014). Do statistical segmentation abilities predict lexical–phonological and lexical–semantic abilities in children with and without SLI? *Journal of Child Language, 41*(2), 327–351. doi:10.1017/S0305000912000736

Mainela-Arnold, E., Evans, J., & Coady, J. A. (2010). Explaining lexical–semantic deficits in specific language impairment: The role of phonological similarity, phonological working memory, and lexical competition. *Journal of Speech, Language, and Hearing Research, 53*(6), 1742–1756. doi:10.1044%2F1092-4388(2010%2F08-0198)

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.

Marton, K., Eichorn, N., Campanelli, L., & Zakariás, L. (2016). Working memory and interference control in children with specific language impairment. *Language and Linguistics Compass, 10*(5), 211–224. doi:10.1111/lnc3.12189

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

*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. doi:10.1080/09297049.2012.734293

McArthur, G., Hogben, J., Edwards, V., Heath, S., & Mengler, E. (2000). On the "specifics" of specific reading disability and specific language impairment. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 41*(7), 869–874. doi:10.1111/1469-7610.00674

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, England: Blackwell.

Meulemans, T., van der Linden, M., & Perruchet, P. (1998). Implicit sequence learning in children. *Journal of Experimental Child Psychology, 69*(3), 199–221. doi:10.1006/jecp.1998.2442

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

Miller, C., Kail, R., Leonard, L., & Tomblin, J. (2001). Speed of processing in children with specific language impairment. *Journal of Speech, Language and Hearing Research, 44*(2), 416–433. doi:10.1044/10924388(2001/034)

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

Misyak, J., & Christiansen, M. (2012). Statistical learning and language: An individual differences study. *Language Learning, 62*(1), 302–331. doi:j.1467-9922. 2010.00626.x

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

Moher, D., Liberati, A., Tetzlaff, J., Altman, D., & The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *Annals of Internal Medicine, 151*(4), 264–269. doi:10.7326/0003-4819-151-4-200908180-00135

Monaghan, P., & Rebuschat, P. (Eds). (2019). Aligning implicit learning and statistical learning: two approaches, one phenomenon [Special issue]. *Topics in Cognitive Neuroscience, 11*(3), 459–467. doi:10.1111/tops.12438

Montgomery, J. (2003). Working memory and comprehension in children with specific language impairment: What we know so far. *Journal of Communication Disorders, 36*, 221–231. doi:10.1016/S0021-9924(03)00021-2

Montgomery, J., Evans, J., & Gillam, R. (2018). Memory and language in children with SLI. In: Alloway, T. (Ed.), *Working memory and clinical developmental disorders.* London and New York: Routledge Taylor & Francis Group.

Montgomery, J., Magimairaj, B. & Finney, M. (2010). Working memory and specific language impairment: An update on the relation and perspectives on assessment and treatment. *American Journal of Speech-Language Pathology, 19*(1), 78–94. doi:10.1044/1058-0360(2009/09-0028)

Newport, E., & Aslin, R. (2004). Learning at a distance I: Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127–162. doi:10.1016/S0010-0285(03)00128-2

Nicolson, R., & Fawcett, A. (2007). Procedural learning difficulties: Reuniting the developmental disorders? *Trends in Neurosciences, 30*(4), 135–141. doi:10.1016/j.tins.2007.02.003

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

Noonan, N. (2018). *Exploring the process of statistical language learning.* (Doctoral dissertation, The University of Western Ontario, London, Canada). Retrieved from: Electronic Thesis and Dissertation Repository (5638)

****Obeid, R. (2017). *Exploring the relationship between sequence learning, motor coordination, and language development* (Doctoral dissertation, The City University of New York, New York, NY).

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. doi:10.3389/fpsyg.2016.01245

Onnis, L., Monaghan, P., Christiansen, M., & 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.

Orpella, J., Ripollés, P., Ruzzoli, M., Amengual, J., Callejas, A., Martines-Alvarez, A., … de Diego-Balaguer, R. (2019). Integrating when and what information in the left parietal lobe allows language rule generalization. Manuscript submitted for publication. Available via: https://www.biorxiv.org/content/10.1101/747816v1

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

Pavlidou, E., & Williams, J. (2014). Implicit learning and reading: Insights from typical children and children with developmental dyslexia using the artificial grammar learning (AGL) paradigm. *Journal in Developmental Disabilities, 35*(7), 1457–1472. doi:10.1016/j.ridd.2014.03.040

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10*(5), 233–238. doi:10.1016/j.tics.2006.03.006

Pinker, S. (1994). *The language instinct.* New York, US: William Morrow & Co.

Plante, E., & Gómez, R. (2018). Learning without trying: The clinical relevance of statistical learning. *Language, Speech, and Hearing Services in Schools, 49*, 710–722. doi: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. doi:10.1044/2014_AJSLP-13-0038

Pullum, G., & Scholz, B. (2002). Empirical assessment of the stimulus poverty arguments. *The Linguistic Review,* 19, 9–50.

Qi, Z, Sanchez Araujo, Y., Georgan, W., Gabrieli, J., & Arciuli, J. (2019). Hearing matters more than seeing: a cross-modality study of statistical learning and reading ability. *Scientific Studies of Reading, 23*(1), 101–115. doi:10.1080/10888438.2018.1485680

Racsmány, M., Lukács, Á., Németh, D., & Pléh, C. (2005). A verbális munkamemória magyar nyelvű vizsgálóeljárásai [Hungarian methods for studying verbal working memory]. *Magyar Pszichológiai Szemle, 60,* 479–505.

Raven, J., Raven, J.C., & Court, J.H. (2003). Manual for Raven's progressive matrices and vocabulary scales. San Antonia: Harcourt Assessment.

R Core Team (2017, 2018). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R foundation for statistical computing. Retrieved from https://www.r-project.org/

Reali, F., & Christiansen, M. (2005). Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cognitive Science, 29*(6), 1007–1028. doi:10.1207/s15516709cog0000_28

Renfrew, C. (2003). The Action Picture Test (4th ed.) [Measurement instrument]. Oxford, England: Speechmark.

Rice, M., Wexler, K., & Cleave, P. (1995). Specific language impairment as a period of extended optional infinitive. *Journal of Speech, Language and Hearing Research, 38*(4), 850–863. doi:10.1044/jshr.3804.850

Rohrmeier, M., Zuidema. W., Wiggins, G., & Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society B, 370*:20140097. doi:10.1098/rstb.2014.0097

Roid, M., & Miller, L. (1997). Leiter International Performance Scale–Revised [Measurement instrument]. New York, NY: Springer.

Romberg, A., & Saffran, J. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906–914. doi:10.1002/wcs.78

Romberg, A., & Saffran, J. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science, 37*(7), 1290–1320. doi:10.1111/cogs.12050

Saffran, J. (2002). Constraints on statistical learning. *Journal of Memory and Language, 47,* 172–196. doi:10.1006/jmla.2001.2839

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

Saffran, J., & Kirkham, N. (2018). Infant statistical learning. *Annual Review of Psychology, 69*, 181–203. doi:10.1146/annurev-psych-122216-011805

Saffran, J., Newport, E., Aslin, R., Tunick, R., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*(2), 101–105. doi:10.1111/j.1467-9280.1997.tb00690.x

Sandoval, M., & Gómez, R. (2013). The development of non-adjacent dependency learning in natural and artificial languages. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(5), 511–522. doi:10.1002/wcs.1244

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

Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia, 67*(2), 147–162. doi:10.1007/s11881-016-0136-0

Schmalz, X., Moll, K., Mulatti, C., & Schulte-Körne, G. (2018). Is statistical learning related to reading ability, and if so, why? *Scientific Studies of Reading, 23*(1)*,* 64–76. doi:10.1080/10888438.2018.1482304

Schwarzer, G. (2015). *Meta: General package for meta-analysis. R package version 4.3-2*. [Statistical software] Retrieved from https://CRAN.R-project.org/package=meta

Semel, E., Wiig, E., & Secord, W. (1987). *CELF-R:* Clinical Evaluation of Language Fundamentals–Revised [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E., & Secord, W. (1995). Clinical Evaluation of Language Fundamentals–Third Edition [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E., & Secord, W. (2003). Clinical Evaluation of Language Fundamentals–Fourth Edition [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

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

Shafto, C., Conway, C., Field, S., & Houston, D. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy*, *17*(3), 247–271. doi:10.1111/j.1532-7078.2011.00085.x

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

Siegelman, N., Bogaerts, L., Elazar, E., Arciuli, J., Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198–213. doi:10.1016/j.cognition.2018.04.011

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

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science, 42*, 1–36. doi:10.1111/cogs.12556

Sigurdardottir, H., Danielsdottir, H., Gudmundsdottir, M., Hjartarson, K., Throrarinsdottir, E., & Kristjánsson, Á. (2017). Problems with visual statistical learning in developmental dyslexia. *Scientific Reports, 7*:606. doi:10.1038/s41598-017-00554-5

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. doi:10.1177/0956797611417632

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

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

Spencer, M., Kaschak, M., Jones, J., & Lonigan, C. (2015). Statistical learning is related to early literacy-related skills. *Reading and Writing: An Interdisciplinary Journal, 28*, 467–490. doi:10.1007/s11145-014–9533-0

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

Steacy, L., Compton, D., Petscher, Y., Elliott, J., Smith, K., Rueckl, J., Sawi, O., Frost, S., & Pugh, K. (2019). Development and prediction of context-dependent vowel pronunciation in elementary readers. *Scientific Studies of Reading, 23*(1), 49–63. doi:10.1080/10888438.2018.1466303

Stichting Siméa (2014). Indicatiecriteria auditief en/of communicatief beperkte leerlingen. Retrieved from: http://www.simea.nl

Tallal, P. (2000). Experimental studies of language learning impairments: From research to remediation. In D. Bishop, & L. Leonard (Eds.), *Speech and language impairment in children* (pp. 131–155). Hove, England: Psychology Press.

Tallal, P., Stark, R., & Mellits, D. (1985). The relationship between auditory temporal analysis and receptive language development: Evidence from studies of developmental language disorder. *Neuropsychologia, 23*(4), 527–534. doi:10.1016/0028-3932(85)90006-5

Thomas, K., & Nelson, C. (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

Thompson, S., & Newport, E. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3*(1), 1–42. doi:10.1207/s15473341lld0301_1

Tomblin, J., Freese, P., & Records, N. (1992). Diagnosing specific language impairment in adults for the purpose of pedigree analysis. *Journal of*

Speech and Hearing Research, 35(4), 832–843. doi:10.1044/jshr.3504.832

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

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

Turk-Browne, N., Jungé, J., & Scholl, B. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552–564. doi:10.1037/0096-3445.134.4.552

Ullman, M. (2015). The declarative/procedural model: A neurobiologically-motivated theory of first and second language. In B. Van Patten & J. Williams (Eds.), *Theories in second language acquisition* (second edition. pp. 135–158). New York and London: Routledge Taylor & Francis Group.

Ullman, M., Earle, F., Walenski, M., & Janacsek, K. (2020) The neurocognition of developmental disorders of language. *Annual Review of Psychology, 71*(5), 5.1–5.29. doi:10.1146/annurev-psych-122216-011555

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

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

Vakil, E., Lowe, M., & Goldfus, C. (2015). Performance of children with developmental dyslexia on two skill learning tasks—serial reaction time and Tower of Hanoi puzzle: A test of the specific procedural learning difficulties theory. *Journal of Learning Disabilities, 48*(5), 471–481. doi:10.1177/0022219413508981

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

van der Kleij, S., Groen, M., Segers, E., & Verhoeven, L. (2018). Sequential implicit learning ability predicts growth in reading skills in typical readers and children with dyslexia. *Scientific Studies of Reading, 23*(1), 77–88. doi:10.1080/10888438.2018.1491582

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

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019a). Statistical learning abilities of children with developmental dyslexia across three experimental paradigms. *PLoS ONE*, *14*(8):e0220041. doi:10.1371/journal.pone.0220041

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019b, 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. doi:10.13140/RG.2.2.24037.96483

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (submitted). Grammatical difficulties in children with dyslexia: The contributions of individual differences in phonological memory and statistical learning.

van Witteloostuijn, M., Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Assessing visual statistical learning in early-school-aged children: The usefulness of an online reaction time measure. *Frontiers in Psychology, 10*:2051. doi:10.3389/fpsyg.2019.02051

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

*von Koss Torkildsen, J. (2010, November). *Event-related potential correlates of artificial grammar learning in preschool children with specific language impairment and controls*. Poster session presented at the Second Annual Neurobiology of Language Conference, San Diego, CA.

von Koss Torkildsen, J., Arciuli, J., & Wie, O. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and individual differences, 69*, 60–68, doi:10.1016/j.lindif.2018.11.003

von Koss Torkildsen, J., Dailey, N., Aguilar, J., Gómez, R., & Plante, E. (2013). Exemplar variability facilitates rapid learning of an otherwise unlearnable grammar by individuals with language-based learning disability. *Journal of Speech, Language, and Hearing Research, 56*(2), 618–629. doi:10.1044/1092-4388(2012/11-0125)

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

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%2F1745691612463078

Wallace, G., & Hammill, D. (1997). The Comprehensive Receptive and Expressive Vocabulary Test: Adult [Measurement instrument]. Austin, TX: Pro-Ed.

Wechsler, D. (1974). Wechsler Intelligence Scale for Children–Revised [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). The Wechsler Intelligence Scale for Children–Third Edition [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). The Wechsler Intelligence Scale for Children–Fourth Edition [Measurement instrument]. San Antonio, TX: The Psychological Corporation.

***West, G., Shanks, D., & Hulme, C. (2018, September 26). Sustained attention, not procedural learning, is a predictor of language, reading and arithmetic skills in children. Retrieved via: doi:10.31234/osf.io/afrms

***West, G., Vadillo, M., Shanks, D., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science, 21*(2):e12552. doi:10.1111/desc.12552

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*. Advance online publication. doi:10.1111/tops.12381

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

Zwart, F., Vissers, C., Kessels, R., & Maes, J. (2019). Procedural learning across the lifespan: A systematic review with implications for atypical development. *Journal of Neuropsychology, 13*(2), 149–182. doi:10.1111/jnp.12139

Appendices

Chapter 2

A2.1: Syntax search term

A2.2: Formulae used

Chapter 4

A4: Operationalization of the model predictors

Chapter 5

A5.1: Visual statistical learning task instructions

A5.2: Visual statistical learning triplets

A5.3: Offline test items visual statistical learning task

Chapter 6

A6.1: Syntax search term

A6.2: Formulae used

Appendices Chapter 2

A2.1: Details of all key words, Boolean operators and syntax used for the database searches

Search strategy for Psychinfo and Eric

((Specific language impairment or Specific Language Disorder or Speech disorder or Communication disorder or Communication delay or Communication impairment or Developmental language delay or Developmental language disorder or Developmental language impairment or Expressive language disorder or Language delay or Language disorder or Language impairment or mixed language disorder or receptive language disorder or Language-based learning disabilit* or Language Based Learning Disabilit* or Language disabled or Specific learning disorder or Learning disabilit*) and (Non-declarative learn* or Non declarative learn* or Nondeclarative learning or Procedural learn* or Implicit learn* or Procedural memory or Artificial grammar learn* or Artificial Grammar or Artificial language or Artificial synta* or Sequence learning or Sequenc* learn* or Statistical learn* or Probabilistic learn* or Non adjacent dependency learning or non adjacent dependencies or nonadjacent dependency learning or nonadjacent dependencies or adjacent dependency learn* or adjacent dependencies)).a

Search strategy for Pubmed

((“Specific language impairment”[Title/Abstract] OR “Specific Language Disorder”[Title/Abstract] OR “Speech disorder”[Title/Abstract] OR “Communication disorder”[Title/Abstract] OR “Communication delay”[Title/Abstract] OR “Communication impairment”[Title/Abstract] OR “Developmental language delay”[Title/Abstract] OR “Developmental language disorder”[Title/Abstract] OR “Developmental language impairment”[Title/Abstract] OR “Expressive language disorder”[Title/Abstract] OR “Language delay”[Title/Abstract] OR “Language disorder”[Title/Abstract] OR “Language impairment”[Title/Abstract] OR “mixed language disorder”[Title/Abstract] OR “receptive language disorder”[Title/Abstract] OR “Language-based learning disabilit*”[Title/Abstract] OR “Language Based Learning Disabilit*”[Title/Abstract] OR “Language disabled”[Title/Abstract] OR “Specific learning disorder”[Title/Abstract] OR “Learning disabilit*”[Title/Abstract])) AND (“Non-declarative learn*”[Title/Abstract] OR “Non declarative learn*”[Title/Abstract] OR “Nondeclarative learning”[Title/Abstract] OR “Procedural learn*”[Title/Abstract] OR “Implicit learn*”[Title/Abstract] OR “Procedural memory”[Title/Abstract] OR “Artificial grammar learn*”[Title/Abstract] OR “Artificial Grammar”[Title/Abstract] OR “Artificial language”[Title/Abstract] OR “Artificial synta*”[Title/Abstract] OR “Sequence learning”[Title/Abstract] OR “Sequenc* learn*”[Title/Abstract] OR “Statistical learn*”[Title/Abstract] OR “Probabilistic learn*”[Title/Abstract] OR “Non adjacent dependency learning”[Title/Abstract] OR “non adjacent dependencies”[Title/Abstract] OR “nonadjacent dependency learning”[Title/Abstract] OR “nonadjacent dependencies”[Title/Abstract] OR “adjacent dependency learn*”[Title/Abstract] OR “adjacent dependencies”[Title/Abstract])

Search strategy for Language Behavioural abstracts (LLB)

ab(Specific language impairment OR Specific Language Disorder OR Speech disorder OR Communication disorder OR Communication delay OR Communication impairment OR Developmental language delay OR Developmental language disorder OR Developmental language impairment OR Expressive language disorder OR Language delay OR Language disorder OR Language impairment OR mixed language disorder OR receptive language disorder OR Language-based learning disability OR Language Based Learning Disability OR Language disabled OR Specific learning disorder OR Learning disability) AND ab(Non-declarative learning OR Non declarative learning OR Procedural learning OR Implicit learning OR Procedural memory OR Artificial grammar learning OR Artificial Grammar OR Artificial language OR Artificial syntax OR Sequence learning OR Sequenced learning OR Statistical learning OR Probabilistic learning OR Non adjacent dependency learning OR non adjacent dependencies OR nonadjacent dependency learning OR nonadjacent dependencies OR adjacent dependency learn* OR adjacent dependencies)

Search strategy for Open Access Thesis and Dissertations (OATD)

abstract:(("Non-declarative learning" OR "Non declarative learning" OR "Procedural learning" OR "Implicit learning" OR "Procedural memory" OR "Artificial grammar learning" OR "Artificial Grammar" OR "Artificial language" OR "Artificial syntax" OR "Sequence learning" OR "Sequenced learning" OR "Statistical learning" OR "Probabilistic learning" OR "Non adjacent dependency learning" OR "non adjacent dependencies" OR "nonadjacent dependency learning" OR "nonadjacent dependencies" OR "adjacent dependency learning" OR "adjacent dependencies") AND abstract:("Specific language impairment" OR "Specific Language Disorder" OR "Speech disorder" OR "Communication disorder" OR "Communication delay" OR "Communication impairment" OR "Developmental language delay" OR "Developmental language disorder" OR "Developmental language impairment" OR "Expressive language disorder" OR "Language delay" OR "Language disorder" OR "Language impairment" OR "mixed language disorder" OR "receptive language disorder" OR "Language-based learning disability" OR "Language Based Learning Disability" OR "Language disabled" OR "Specific learning disorder" OR "Learning disability"))

A2.2: Formulae used

	A) Independent groups: Means and SD for SLI and Controls available (N= 7)	**B)** Independent groups: t-statistic or F-statistic (between subjects) available (N= 3)
Step 1 effect size calculation	$SMD = \frac{\overline{X}_{\mathrm{SLI}} - \overline{X}_{\mathrm{Control}}}{SD_{\mathrm{pooled}}}$ $SD_{pooled} = \sqrt{\frac{(N_{\mathrm{SLI}} - 1)SD_{\mathrm{SLI}}^{2} + (N_{\mathrm{Control}} - 1)SD_{\mathrm{Control}}^{2}}{N_{\mathrm{SLI}} + N_{\mathrm{Control}} - 2}}$	$SMD = t\sqrt{\frac{(N_{SLI} + N_{Control})}{N_{SLI}N_{Control}}}$ $SMD = \sqrt{\frac{F(N_{SLI} + N_{Control})}{N_{SLI}N_{Control}}}$
Step 2 Variance of SMD	$\mathrm{var}(SMD) = \frac{N_{SLI} + N_{Control}}{N_{SLI}N_{Control}} + \frac{SMD^{2}}{2(N_{SLI} + N_{Control})}$	
Step 3 Hedges's *g* *correction for* small sample size	$J = 1 - \frac{3}{4df - 1}$ $g = J * SMD$ $\mathrm{var}(g) = J^{2} * \mathrm{var}(SMD)$	
Step 4 Variance of effect size and inverse variance weights	$Weight = \frac{1}{\mathrm{var}(g)}$	

Appendix Chapter 4
A4: Operationalization of the model predictors

Predictor	Sum-to-zero contrasts	Operationalization
Intercept (online measure)		Average normalized response time
DisruptionPeak (online measure)	Disruption block: $+\frac{2}{3}$ Third training block: $-\frac{1}{3}$ Recovery block: $-\frac{1}{3}$	Difference in normalized response time between disruption block and combined third training block and recovery block
PrePostDisruption (online measure)	Disruption block: 0 Third training block: $-\frac{1}{2}$ Recovery block: $+\frac{1}{2}$	Difference in normalized response time between third training block and recovery block
Targetness (online measure)	NonTarget: $-\frac{1}{2}$ Target: $+\frac{1}{2}$	Difference in normalized response times between nontargets and targets
ExpVersion (online + offline measures)	Version 1: $-\frac{1}{2}$ Version 2: $+\frac{1}{2}$	Difference in normalized response times/odd ratio between experiment version 1 (target = *lut*) and experiment version 2 (target = *mip*)
Group (online + offline measures)	DLD: $-\frac{1}{2}$ TD: $+\frac{1}{2}$	Difference in normalized response time/odds ratui between children with DLD and typically developing children
Intercept (offline measure)		Yes bias: difference in odds ratio between children's yes responses and their no responses.

(*Table continues*)

Operationalization of the model predictors (*Continued*)

Rule (offline measure)	NAD rule: $+\frac{1}{2}$ Violation rule: $-\frac{1}{2}$	Difference in odds ratio between items that follow the rule and items that violate the rule
Generalization (offline measure)	Familiar: $+\frac{1}{2}$ Novel: $-\frac{1}{2}$	Difference in odds ratio between items with familiar X-elements and items with novel X-elements

Note. DLD = developmental language disorder. TD = typically developing.

Appendices Chapter 5

A5.1: Visual statistical learning task instructions

Instructions online self-paced familiarization phase

Dutch original: In dit spel staan aliens in de rij voor het ruimteschip. Ze willen graag naar huis. Kan jij helpen? Je ziet straks alle aliens die in de rij staan. Je ziet steeds één alien tegelijk. Stuur de alien naar huis door op de spatiebalk te drukken. Daarna zie je vanzelf de volgende alien in de rij. Probeer maar!

English translation: In this game, aliens are lined up in front of the spaceship. They all want to go home, and it's your task to help them. You will see all of the aliens standing in the line. You will see one alien at a time. Send the alien home by pressing the space bar. After pressing the space bar, you will automatically see the next alien standing in the line. Give it a try!

Dutch original: Goed zo! Dat is makkelijk hè? In dit spel vinden sommige aliens elkaar heel leuk. Zij staan bij elkaar in de rij! Bekijk elke alien goed en kijk welke aliens bij elkaar in de rij staan. Ik stel je hier later nog wat vragen over, dus let heel goed op! We gaan even oefenen.

English translation: Well done! Easy, isn't it? In this game, some aliens really like each other. They stand together in line. Watch each alien closely and pay attention to the order of the aliens, because I will ask you some questions about this later on. We start with a practice.

Dutch original: Goed gedaan! Soms zie je in dit spel dezelfde alien twee keer achter elkaar. Als je dat ziet, moet je de alien wegjagen. Dit doe je door hem aan te raken. Je kan gewoon met je vinger op het scherm drukken.

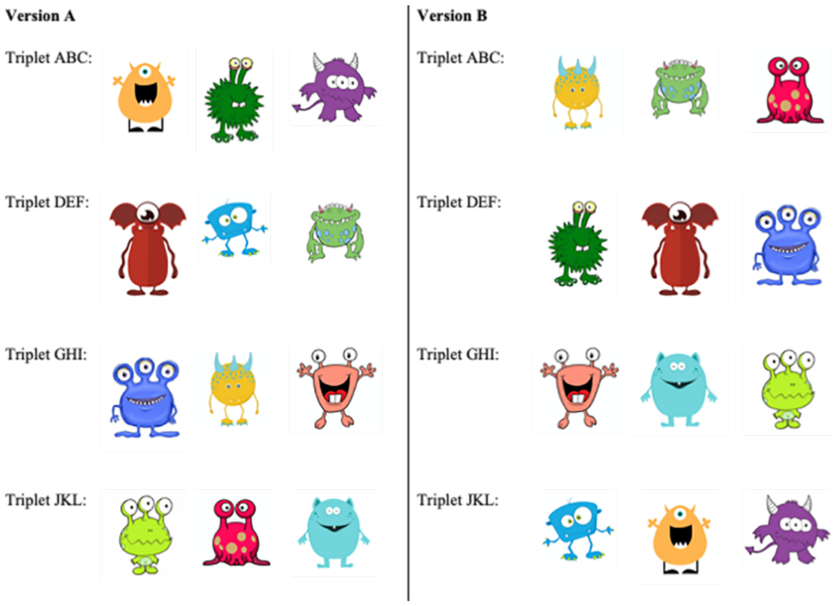
English translation: Well done! In this game, sometimes the exact same alien appears two times in a row. If you see the exact same alien twice in a row, you have to scare the alien away. You can do this by touching him on the screen with your finger.

Dutch original: Dat ging goed! Ben je klaar om echt te beginnen? Vergeet niet om goed op de aliens te letten. Bekijk elke alien goed en kijk welke aliens bij elkaar in de rij staan. Hierover krijg je later nog wat vragen, dus let heel goed op! Als je een alien twee keer achter elkaar ziet, jaag hem dan weg! Daar gaan we, zet hem op!
English translation: Well done! Are you ready for the real game? Don't forget to watch each alien closely and to pay attention to the order of the aliens, because I will ask you some questions about this later on. Also, if you see the exact same alien twice in a row, scare the repeated alien away. Let's go!

Instructions for the offline pattern completion task
Dutch original: Nu gaan we nog iets anders doen. Sommige aliens vonden elkaar heel leuk en stonden daarom bij elkaar in de rij. Als het goed is, heb jij hierop gelet! Daar krijg je nu een paar vragen over. Je ziet steeds bovenaan een plaatje met aliens die steeds bij elkaar stonden, maar… één van de aliens is weg! Jij moet kiezen welke alien op de plek van het vraagteken hoort. Je mag één van de drie aliens kiezen die onderaan staan. Welke alien stond steeds op de plek van het vraagteken? Als je het niet zeker weet, mag je raden.
English translation: Now, we are up for something different. Some aliens really liked each other and stood in line together. Did you pay attention to this? You will now receive some questions about his. On the top of the screen, you will see a picture of aliens that stood together in line, but there is one missing alien {the missing alien is depicted by a question mark}. You have to decide which alien should replace the question mark. You may choose one of the three aliens that have appeared on the bottom of the screen. If you don't know the answer, you may guess.

Instructions for the Pattern recognition task
Dutch original: Je ziet steeds twee plaatjes. Op allebei de plaatjes staat een groepje aliens. Een van deze plaatjes klopt: deze aliens stonden steeds bij elkaar in de rij, in dezelfde volgorde. Jij moet kiezen welke van de twee plaatjes klopt. Als je het niet zeker weet, mag je raden.
English translation: Now, you will see pictures with two groups of aliens. One of the groups of aliens is correct: these aliens stood together in line, in the same order. You may decide which of the two groups of aliens is correct. If you don't know the answer, you may guess.

A5.2: Visual statistical learning triplets



A5.3: Offline test items visual statistical learning task

Overview of the test items (for order 2 of the experiment version with Triplets A and Triplets B). The correct answers are underscored and in bold. Each letter represents an individual alien. The question mark indicates the missing alien.

Triplet completion task

Item	*Triplet/pair to complete*	*Answer options*
1	?C	**B**DK
2	?K	I**J**A
3	GH?	D**I**L
4	D?F	G**E**B
5	?KL	**J**CE
6	?BC	GH**A**
7	?HI	**G**LA
8	K?	H**L**I
9	?E	JK**D**
10	B?	F**C**E
11	A?C	J**B**H
12	G?	KA**H**
13	E?	**F**GC
14	H?	**I**DL
15	J?L	EF**K**
16	DE?	CB**F**

Triplet recognition task

Item	*Triplets/pairs presented on the left side*	*Triplets/pairs presented on the right side*
1	**DEF**	JBF
2	BF	**EF**
3	GK	**HI**
4	KC	**DE**
5	**EF**	GK
6	**ABC**	DHL
7	JBF	GHI
8	**AB**	DH
9	**JKL**	AEI

(*Overview offline test items continues*)

Overview offline test items (*continued*)

Triplet recognition task		
Items	*Triplets/pairs presented on the left side*	*Triplets/pairs presented on the right side*
10	GKC	**ABC**
11	EI	**BC**
12	**JK**	HL
13	KC	**KL**
14	**HI**	HL
15	**GHI**	AEI
16	**BC**	BF
17	**DE**	AE
18	JB	**AB**
19	**GH**	EI
20	**KL**	AE
21	DH	**GH**
22	JB	**JK**
23	GKC	**DEF**
24	DHL	**JKL**

Appendices Chapter 6

A6.1: Details of all key words, Boolean operators and syntax used for the database searches

Note that these are the search strategies of our latest search (January 2019)

Search strategy for Psychinfo and Eric

[Field 1]: (Children) OR (school) OR (*school*)

AND [Field 2]: (Non-declarative learn*) OR (Non declarative learn*) OR (Nondeclarative learning) OR (Procedural learn*) OR (Implicit learn*) OR (Procedural memory) OR (Serial Reaction Time) OR (Serial Reaction Time Task) OR (Sequence learning) OR (Sequenc* learn*) OR (Statistical learn*) OR (Probabilistic learn*)

AND [Field 3]: (Grammar) OR (Grammatical skills) OR (Grammar*) OR (Grammatical abilities) OR (Grammatical abilit*) OR (Language skills) OR (Language skill) OR (Language) OR (Language abilit*) OR (Sentence repetition) OR (Sentence-picture match*) OR (Sentence picture match*) OR (TROG*) OR (Morphosynta*) OR (Morphosyntactic comprehension) OR (Morphology) OR (Morphological skills) OR (Morphological abilit*) OR (Sentence completion) OR (Sentence comprehension) OR (past tense) OR (past-tense) OR (Sentence production) OR (Action Picture Naming) OR (Picture naming)

Search strategy for Pubmed

[Field 1]: “children” OR “school” OR “*school*”

AND [Field 2]: “Non-declarative learn*” OR “Non declarative learn*” OR “Nondeclarative learning” OR “Procedural learn*” OR “Implicit learn*” OR “Procedural memory” OR “Serial Reaction Time” OR “Serial Reaction Time Task” OR “Sequence learning” OR “Sequenc* learn*” OR “Statistical learn*” OR “Probabilistic learn*”

AND [Field 3]: “Grammar” OR “Grammatical skills” OR “Grammar*” OR “Grammatical abilities” OR “Grammatical abilit*” OR “Language skills” OR “Language skill” OR “Language” OR “Language abilit*” OR “Sentence repetition” OR “Sentence-picture match*” OR “Sentence picture match*” OR “TROG*” OR “Morphosynta*” OR “Morphosyntactic comprehension” OR “Morphology” OR “Morphological skills” OR “Morphological abilit*” OR “Sentence completion” OR “Sentence comprehension” OR “past tense” OR “past-tense” OR “Sentence production” OR “Action Picture Naming” OR “Picture naming”

Search strategy for Linguistics and language behavioral abstracts

[Field 1]: children OR school OR *school*

AND [Field 2]: Non-declarative learn* OR Non declarative learn* OR Nondeclarative learning OR Procedural learn* OR Implicit learn* OR Procedural memory OR Serial Reaction Time OR Serial Reaction Time Task OR Sequence learning OR Sequenc* learn* OR Statistical learn* OR Probabilistic learn*

AND [Field 3]: Grammar OR Grammatical skills OR Grammar* OR Grammatical abilities OR Grammatical abilit* OR Language skills OR Language skill OR Language OR Language abilit* OR Sentence repetition OR Sentence-picture match* OR Sentence picture match* OR TROG* OR Morphosynta* OR Morphosyntactic comprehension OR Morphology OR Morphological skills OR Morphological abilit* OR Sentence completion OR Sentence comprehension OR past tense OR past-tense OR Sentence production OR Action Picture Naming OR Picture naming

A6.2: Formulae used

Formulas used to compute effect size information

A. Pearson's *r* into Fisher's *z:*

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

B. Variance Fisher's *z*

$$V_z = \frac{1}{n-3}$$

C. Kendaull's tau (τ) into Pearson's *r*

$$r = \sin\left(\frac{1}{2}\pi\tau\right)$$

D. Synthesized effect size

$$Y_{syn} = \frac{1}{2}(Y_1 + Y_2)$$

Y_1 is outcome 1 (Pearson's *r* between serial reaction time task index and expressive grammar index 1) and Y_2 is outcome 2 (Pearson's *r* between serial reaction time task index and expressive grammar index 2).

E. Synthesized variance

$$V_y = \frac{1}{4}\left(V_{y1} + V_{y2} + 2r\sqrt{V_{Y1}} + \sqrt{V_{Y2}}\right)$$

Where *r* is the Pearson correlation between children's expressive grammar index 1 and expressive grammar index 2.

Summary in English

Detecting patterns: Relating statistical learning to language proficiency in children with and without developmental language disorder

Children differ in the (apparent) ease with which they acquire the sounds, word meanings and rules of their native language. Some children have so many difficulties acquiring language that it has significant impact on their social interactions and educational process. If there is no clear aetiology (e.g., hearing impairment, neurological deficit, deprivation of linguistic input or limited cognitive abilities) for the language difficulties observed, then a child is usually diagnosed with developmental language disorder (DLD). Despite large heterogeneity in the language difficulties observed among children with DLD, almost all children with DLD have problems acquiring grammar of their native language. Grammatical problems are therefore regarded as a clinical marker of the disorder. In addition to their language problems, children with DLD often exhibit deficits in other cognitive areas, such as verbal working memory, verbal short-term memory and attention as well. It is still an empirical question how the language problems in these children can be explained: are they the consequence of a language-specific deficit or do they result from deficits in other cognitive mechanisms that are presumably important for language learning? The research described in this book addresses the latter question. Specifically, the overall aim is to determine if the language problems observed in children with DLD may be the consequence of these children being less sensitive to rules, patterns and regularities in the environment than their typically developing peers.

Statistical patterns and distributions in (spoken) language reflect underlying phonological, morphological and syntactic structures. For example, in the English present tense, the third person pronoun *he* frequently co-occurs with verb-plus-[s] marking as in *he eats, he talks, he walks.* Children unconsciously detect such co-occurrences, which guides them in learning the “rules” or “patterns” of their language. Reduced sensitivity to such statistical regularities in the input may hinder the detection of rules and patterns, and as such a *statistical learning deficit* has been proposed to explain the difficulties that children with DLD have in acquiring the grammar of their language. The example above may suggest that the statistical learning mechanism concerns the detection of *linguistic*

regularities specifically. The latter is not necessarily the case, however. Structure is not unique to human language: other domains (such as music, bird song, or movement) are also organized in a structured way. Therefore, it has been hypothesized that humans have a statistical learning mechanism that operates independently of modality (visual, auditory, visuomotor) and domain (verbal, nonverbal). Therefore, people's sensitivity to regularities in general may correlate with their language proficiency. That is, people who are relatively good at detecting all sorts of regularities in their environment are expected to have relatively high language proficiency.

The present dissertation has three aims: first, the studies described in this booked aimed to contribute to the methodological debate on how to measure statistical learning. The second aim is to investigate whether individual differences in statistical learning correlate with language proficiency. Third, we investigated whether children with DLD have a general statistical learning deficit that may contribute to the language problems observed in these children. In what follows, a summary of our findings with respect to each of these three aims is provided. We start discussing the third aim.

If children with DLD have a general statistical learning deficit, then one would expect to observe reduced sensitivity to regularities in these children as compared to their typically developing peers across domains and modalities. We compared the statistical learning ability of 37 children with DLD and 37 typically developing children (8-12 years old) on three different statistical learning tasks: an auditory verbal task (Chapter 4), a visual nonverbal task (Chapter 5) and a visuomotoric nonverbal task (Chapter 6).

In the auditory verbal task (Chapter 4), we presented children with three-word utterances of a miniature artificial language (e.g. *tep wadim lut, sot kasi mip*). Unbeknownst to the children, the first word of each utterance "predicted" the third word of the utterance. That is, *tep* and *lut* always went together and *sot* and *mip* always went together. While the children listened to these utterances, we asked them to press a green or red button. Which button colour they were supposed to press depended on the third word of the utterance. For example, we asked them to press the green button if the third word was *lut* and the red button if the third word was not *lut*. After a series of rule-blocks – blocks in which the first word of each utterance predicted the third word of the utterance – children were presented with utterances that did not follow the rules. The third word of these non-rule utterances was still *lut* or *mip*, but the first word had changed (e.g., *pif wadim lut*). We reasoned that if children had learned the rule between the first and third word, then this would be visible in their response time pattern: their response times to the third word of utterances that followed the rule would be quicker than their response times to the third word of utterances that did not follow the rule. After all, if they had detected the dependency between the first word and third word of the utterances, they would be able to predict the third word of rule items upon hearing the first word, whereas such a prediction does not work for non-rule items. Comparing the response time patterns of children with DLD to children without DLD, we observed that typically developing children showed the expected pattern: they responded slower to non-rule items than to rule items. We found no evidence for such a difference in children with DLD. When comparing learning between with DLD and their typically developing peers, we found that the learning effect was smaller in children with DLD than in typically developing children. We concluded that children with DLD have an auditory verbal statistical learning deficit. However, we also observed that the deficit was small in size.

In the visual nonverbal task (Chapter 5), we told children that they were going to play a game in which it was their task to send alien creatures back to their home planet. We instructed children that a couple of aliens stood behind each other in line, waiting to board a spaceship. The children would see one alien at a time and they could send the alien home by pressing the space bar. After they had pressed the space bar, the next alien in line appeared automatically. We also told the children that they had to pay attention to the order in which the aliens appeared, because later on, we would ask them some questions about this order. What the children did not know was that the aliens formed triplets: there were twelve different aliens and these were arranged in four groups of three aliens. Thus alien 1, alien 2 and alien 3 belonged together; alien 4, alien 5 and alien 6 belonged together; alien 7, alien 8 and alien 9 belonged together, and alien 10, alien 11, and alien 12 belonged together. This also meant that alien 1 was always followed by alien 2 and that alien 2 was always followed by alien 3. That is, the transitional probability from alien 1 to alien 2 was 100%. Also, the transitional probability from alien 2 to alien 3 was 100%. However, the transitional probability from alien 3 to the next alien was lower, as alien 3 could be followed by the first alien of any of the three other triplets (i.e. alien 4, alien 7 or alien 10). Previous studies showed that that learners are sensitive to these differences in transitional probabilities and that they use them to distinguish high-probability sequences from low-probability sequences. The latter facilitates the learning of triplets: the transitional probability between aliens that form a triplet is higher than the transitional probability between aliens that cross a triplet boundary. As for the auditory verbal task, we compared the learning of children with DLD to the learning of children without DLD. Interestingly, we observed that both groups of children learned which aliens belonged together. This outcome suggests that children with DLD are able to detect statistical regularities in the visual nonverbal domain.

Finally, we assessed children's sensitivity to sequenced patterns in the visuomotoric nonverbal domain (Chapter 6), using a serial reaction time task. Children were seated in front of a computer screen with a gamepad controller attached to it. A cartoon picture of a smiley appeared in one of four marked locations on the screen. We instructed children to press the corresponding button on the gamepad controller as quickly and accurately as possible. In the first four blocks and the final block of the experiment the appearances of the smiley followed a fixed sequence of 10 screen positions. In the fifth (pre-final) block, the fixed sequence was replaced by a random one. Similarly to the auditory verbal task, children's response time patterns served as an index of learning: if children detect the sequence, then their response times in the sequenced blocks should be quicker than their response times in the random block. We observed the predicted difference in both groups of children. This suggests that both children with and without DLD learned the sequence in the visuomotor nonverbal domain.

Taken together, we can only conclude that children with DLD performed differently from typically developing peers on the auditory verbal statistical learning task. In both non-auditory nonverbal statistical learning tasks we observed that children with DLD did detect the regularities. This pattern of results may suggest that the statistical learning deficit is restricted to the auditory verbal domain. However, more research is needed to confirm that the difference between children with DLD and typically developing children in this domain is indeed a consequence of reduced sensitivity to regularities and not a result of reduced linguistic entrenchment, or deficits in auditory processing in children with DLD as compared to typically developing children.

The summary above described differences in statistical learning performance between children with and without DLD at the group level. We also investigated the correlation between an individual children's statistical learning ability and individual measures of grammatical proficiency, reading proficiency and spelling proficiency. We predicted that children who are relatively good at detecting the patterns, rules and regularities in our statistical learning tasks score also relatively high on our measures of language proficiency. However, we found no evidence for or against this hypothesis. In Chapter 4 we could not detect a correlation between children's learning of the rules in the miniature artificial language and their (native language) grammatical proficiency. In Chapter 5 we found no evidence for or against a correlation between children's learning of the alien triplets and their scores on reading and spelling tasks, and in Chapter 6 we found no evidence for or against a correlation between children's learning of the smiley sequence and their scores on a grammatical proficiency task.

One of the explanations for these inconclusive results on the association between statistical learning and language proficiency is that our measures of statistical learning were not good enough to measure statistical learning at the individual level. We are not the first to suggest this may be at stake. Within the field of statistical learning, other research groups raised their concerns on the sensitivity and reliability of the commonly used measures of statistical learning. Some groups came with recommendations to improve the sensitivity of the measures. The visual statistical learning task that we used in Chapter 5 implemented some of these recommendations.

Another methodological discussion in the statistical learning literature is whether statistical learning should be measured *while* people are learning (online measure) or *after* learning took place (offline measure). There is a rise in statistical learning tasks that use both online and offline measures of statistical learning. However, that the use of online measures is relatively new, is illustrated by the meta-analysis that we conducted at the start of this PhD project (Chapter 2). This overview shows that only one out of the ten included studies on auditory verbal statistical learning that we included in our quantitative overview used an online measure (event-related potentials) of learning. All other studies used only offline measures of learning. This observation was one of the reasons for us to develop a novel child-friendly online measure of auditory verbal statistical learning. In Chapter 3 we showed that this novel measure can be used to detect learning in primary-school-aged children between five and eight years old. In Chapter 4 we continued to use this measure to investigate differences in auditory verbal statistical learning between children with and without DLD (see above). As the use of online measures of statistical learning is relatively new, novel measures keep emerging. In Chapter 5 we therefore used a slightly different online measure of statistical learning. That is, in Chapters 4 and 6, we used the difference in children’s response times to rule items (faster responses) versus nonrule items (slower responses) as an index of learning. In Chapter 5 we used the difference in children’s response times to predictable (faster responses) versus less predictable (slower responses) as an index of learning. However, using this slightly different measure, we could not detect learning in children with and without DLD. This shows that it is still an empirical question what, and under which conditions, is the best online measure of statistical learning.

Having discussed the three main aims of this dissertation, we conclude that the presence of a statistical learning deficit in children with DLD as well as the strength of the correlation between children’s individual statistical learning ability and their language proficiency may depend on several factors, including but not restricted to the domain and modality in which learning is tested and the way in which statistical learning is measured. We concluded that children with DLD have an auditory verbal statistical learning deficit, but could not conclude that they have (or do not have) a statistical learning deficit outside this domain. More research is needed to confirm that the observed difference between children with and without DLD in the auditory verbal domain is indeed the consequence of reduced statistical learning and not of deficiencies in other cognitive areas such as auditory processing or reduced linguistic entrenchment in people with DLD. Though not investigated as such, our results may also be a preliminary indication that within a treatment context, interventions that aim to bolster children’s statistical learning may have limited, if any, effects. Therefore, it may be more effective to focus on the training of other aspects that are more strongly correlated to children’s language proficiency than statistical learning.

Nederlandse samenvatting

Het ontdekken van patronen: De relatie tussen statistisch leren en taalvaardigheid in kinderen met en zonder taalontwikkelingsstoornis

Kinderen verschillen in het (ogenschijnlijke) gemak waarmee zij de klanken, woordbetekenissen en grammaticale regels van hun moedertaal leren. Sommige kinderen hebben zoveel moeite met het leren van taal dat het negatieve gevolgen heeft voor hun sociale interacties en schoolprestaties. Wanneer de problemen met het leren van taal geen duidelijke oorzaak hebben, zoals bijvoorbeeld gehoorverlies, een neurologische afwijking of een sterk verminderd taalaanbod, dan spreken we van een "taalontwikkelingsstoornis" (TOS). Naast problemen met taal hebben veel kinderen met TOS ook problemen met andere cognitieve vaardigheden, zoals werkgeheugen, aandacht en kortetermijngeheugen. Deze cognitieve vaardigheden zijn ook belangrijk voor het leren van taal. Het is daarom een empirische vraag of de taalproblemen bij kinderen met TOS het gevolg zijn van problemen in taalspecifieke leermechanismen of van problemen in bredere cognitieve leermechanismen. Een algemeen cognitief leermechanisme waarvan verondersteld wordt dat het belangrijk is voor taalontwikkeling, is statistisch leren. Een statistisch leermechanisme zou mensen in staat stellen regelmatigheden in hun omgeving te detecteren. Waarom dit belangrijk kan zijn voor taalontwikkeling leggen we hieronder uit.

In iedere taal komen bepaalde elementen, zoals klanken, lettergrepen of morfemen, relatief vaak in bepaalde combinaties voor. Bijvoorbeeld, in het Nederlands worden de meeste werkwoorden (2e en 3e persoon) in de tegenwoordige tijd vervoegd als werkwoordstam + *t* (*loopt, fietst, danst*). Dit betekent dat de overgangswaarschijnlijkheid tussen enkelvoudige onderwerpen (*hij, jij, Anna*) en een werkwoordstam + *t* relatief hoog is (*hij loopt, jij fietst, Anna danst*). Steeds meer onderzoek laat zien dat kinderen gevoelig zijn voor dit soort statistische regelmatigheden en dat deze gevoeligheid kinderen helpt om onbewust de grammaticale regels van hun moedertaal te leren.

Regelmatigheden komen niet alleen voor in menselijke taal. Ook in andere domeinen, zoals muziek, vogelzang en motoriek hebben een (statistische) structuur. Er wordt verondersteld dat mensen een algemeen, cognitief statistisch leermechanisme gebruiken om regelmatigheden in allerlei domeinen te detecteren. Er wordt ook verondersteld dat dit algemene statistisch leermechanisme een rol speelt tijdens het leren van taal: mensen die regelmatigheden in hun omgeving relatief snel oppikken, zouden dan ook een relatief goede taalvaardigheid hebben. Tegelijkertijd betekent dit dat de taalproblemen zoals we die zien bij kinderen met TOS, deels verklaard zouden kunnen worden door een statistisch leerprobleem. Kinderen met TOS zouden minder gevoelig zijn voor regelmatigheden dan kinderen zonder TOS.

Centraal in dit proefschrift staat de hierboven beschreven relatie tussen statistisch leren en taalvaardigheid. In verschillende studies is onderzocht of individuele verschillen in statistisch leren samenhangen met taalvaardigheid én of kinderen met TOS minder gevoelig zijn voor statistische regelmatigheden in hun omgeving dan kinderen zonder TOS. Daarnaast levert dit proefschrift ook een bijdrage aan het methodologisch debat rondom de manier waarop statistisch leren het beste gemeten kan worden.

Als een algemeen statistisch leerprobleem bijdraagt aan de taalproblemen die we zien bij kinderen met TOS, dan verwachten we dat kinderen met TOS zwakker zijn voor het herkennen van regelmatigheden in verschillende soorten aanbod dan kinderen zonder TOS. Om deze verwachting te toetsen, onderzochten wij het statistisch leervermogen van 37 kinderen met TOS en 37 kinderen zonder TOS op drie verschillende statistisch leertaken: een auditieve talige taak (Hoofdstuk 4), een visuele niet-talige taak (Hoofdstuk 5) en een visueelmotorische niet-talige taak (Hoofdstuk 6). Alle kinderen waren tussen de 8 en 12 jaar oud.

In de auditieve talige taak (Hoofdstuk 4) luisterden de kinderen naar een niet bestaande taal. Deze niet bestaande taal bestond uit driewoordzinnen, zoals *tep wadim lut* en *sot kasi mip.* Zonder dat de kinderen het wisten zat er een regel verstopt in de taal: het eerste woord van de zin voorspelde het derde woord van de zin. Dus *tep* voorspelde *lut* en *sot* voorspelde *mip*. Terwijl de kinderen naar de niet bestaande taal luisterden vroegen wij ze op een groene of rode knop te drukken. Welke knop ingedrukt moest worden, was afhankelijk van het derde woord uit de zin. Bijvoorbeeld, kinderen moesten op de groene knop drukken wanneer het derde woordje *lut* was en op de rode knop wanneer het derde woord geen *lut* was. Kinderen luisterden eerst gedurende een aantal blokken naar zinnen die de *tep-lut* en *sot-mip* regel volgden. Vervolgens was er een blok waarin de regel doorbroken werd. Dit blok noemen we het *disruptieblok*. In het disruptieblok eindigden de zinnen nog steeds met *lut* of *mip,* maar was het eerste woord variabel (dus bijvoorbeeld *pif wadim lut* en *gak palti mip*). Het idee hierachter is dat wanneer kinderen de regel geleerd hebben, ze in het disruptieblok langzamer op de groene of rode knop zullen drukken dan in de blokken waar de regel wel aanwezig is. In het disruptieblok voorspelt het eerste woord niet langer het derde woord, wat voor een langere reactietijd zorgt. Als we de patronen in reactietijd tussen kinderen met en zonder TOS vergelijken zien we dat het verschil in reactietijd tussen de regelblokken en het disruptieblok groter is bij kinderen zonder TOS dan bij kinderen met TOS. Kinderen zonder TOS drukken sneller voor zinnen die de regel volgen dan zinnen die de regel niet volgen. Bij kinderen met TOS vinden we geen bewijs voor zo'n verschil in reactietijd tussen regels en niet regels. We concluderen dat kinderen met TOS de regel minder goed hebben geleerd dan kinderen zonder TOS, en dus dat zij mogelijk een auditief talig statistisch leerprobleem hebben.

In Hoofdstuk 5 beschrijven we het statistisch leren van kinderen met en zonder TOS op een visuele niet-talige taak. In deze taak kregen kinderen de opdracht om aliens naar huis te sturen. De aliens stonden achter elkaar in de rij te wachten op een ruimteschip en telkens wanneer het kind op de spatiebalk drukte, ging er een alien naar huis en verscheen de volgende alien uit de rij. We vertelden de kinderen dat ze goed moesten opletten in welke volgorde de aliens verschenen. De kinderen wisten niet dat er twaalf aliens waren, die telkens verschenen in groepjes van drie. Dus alien 1, alien 2 en alien 3 verschenen altijd na elkaar, alien 4, alien 5 en alien 6 verschenen altijd na elkaar, alien 7, alien 8 en alien 9 verschenen altijd na elkaar, en alien 10, alien 11 en alien 12 verschenen altijd na elkaar. We noemen deze groepjes van drie aliens *triplets* en er zijn dus vier alientriplets. Binnen zo'n triplet is de overgangswaarschijnlijk van de ene alien naar de volgende alien 100%, want na alien 1 volgt altijd alien 2 en na alien 2 volgt altijd alien 3. De overgangswaarschijnlijkheid *tussen* triplets is lager, want alien 3 kan gevolgd worden door de eerste alien van een van de drie andere triplets (alien 4, alien 7 of alien 10). Als kinderen gevoelig zijn voor deze verschillen in overgangswaarschijnlijkheid, dan leren zij welke aliens bij elkaar horen en dus een triplet vormen. Aan het einde van de taak bleek dat zowel kinderen met TOS als kinderen zonder TOS de alien triplets herkenden. Dit betekent dat ook kinderen met TOS gevoelig zijn voor regelmatigheden in het visuele niet-talige domein.

In Hoofdstuk 6 onderzochten we statistisch leren in het visueelmotorische niet-talige domein. Een taak die vaak gebruikt wordt om dit type statistisch leren te onderzoeken is de *serial reaction time* taak. Tijdens deze taak verschijnt er herhaaldelijk een *smiley,* maar steeds op andere locaties. Er zijn vier mogelijke locaties op een computerscherm. Kinderen krijgen een *gamepad controller* met daarop vier knoppen die overeenkomen met de vier locaties op het scherm. Het is aan de kinderen om zo snel mogelijk op de knop te drukken die overeenkomt met de locatie op het scherm waar de smiley verschijnt. In de eerste vier blokken en het zesde (laatste) blok van het experiment verschijnt de smiley in een vaste volgorde van 10 locaties. De kinderen weten dit echter niet. In het vijfde (voorlaatste) blok verschijnt de smiley in een willekeurige volgorde. Net als bij de auditieve talige taak (zie boven) verwachten we een verschil in reactietijd voor smileys die volgens het patroon verschijnen ten opzichte van smileys die willekeurig verschijnen. Dit verschil in reactietijd werd gevonden voor zowel kinderen met TOS als kinderen zonder TOS. Alle kinderen drukten sneller op de knoppen wanneer de smiley volgens de vaste volgorde verscheen dan wanneer de smiley willekeurig verscheen. Dit betekent dat ook kinderen met TOS gevoelig zijn voor regelmatigheden in het visueelmotorische niet-talige domein.

Aan de hand van de resultaten uit de Hoofdstukken 4, 5 en 6 concluderen we dat kinderen met TOS een auditief talig statistisch leerprobleem hebben. In het niet-talige domein zijn kinderen met TOS wel gevoelig voor regelmatigheden. In de algemene discussie van dit proefschrift (Hoofdstuk 7) nuanceren we onze conclusie: vervolgonderzoek moet uitwijzen of het verschil tussen kinderen met en zonder TOS op de auditieve talige taak inderdaad het gevolg is van een statistisch leerprobleem of dat het bijvoorbeeld samenhangt met problemen in de verwerking van auditieve of talige stimuli.

De hierboven beschreven resultaten gaan over verschillen in statistisch leren op groepsniveau. Ook op individueel niveau hebben we de samenhang tussen statistisch leren en verschillende vormen van taalvaardigheid bestudeerd. Het was hierbij de verwachting dat kinderen die goed zijn in statistisch leren ook goed scoren op taken die taalvaardigheid meten (en vice versa). Onze resultaten kunnen deze hypothese niet bevestigen of verwerpen. In Hoofdstuk 4 vonden we geen bewijs voor (of tegen) een samenhang tussen het leren van regels in de niet-bestaande taal en grammaticale vaardigheid. In Hoofdstuk 5 vonden we geen bewijs voor (of tegen) een samenhang tussen het herkennen van alien triplets en scores op lees, - en spellingtaken. Tot slot, in Hoofdstuk 6 vonden we geen bewijs voor (of tegen) een samenhang tussen scores op de serial reaction time taak en grammaticale vaardigheid.

Een van de mogelijke verklaringen voor het niet kunnen detecteren van correlaties (samenhang) tussen statistisch leren en taalvaardigheid op individueel niveau, is dat de bestaande statistisch leertaken niet gevoelig genoeg zijn om verschillen op individueel niveau waar te nemen. Binnen het onderzoeksveld van statistisch leren is dit een bekend probleem. De afgelopen jaren zijn er dan ook een aantal aanbevelingen gedaan die de bestaande taken gevoeliger zouden maken in het detecteren van individuele verschillen. In de visuele niet-talige taak die wij gebruikt hebben in Hoofdstuk 5 zijn een aantal van deze aanbevelingen verwerkt.

Een andere methodologische discussie die gevoerd wordt binnen het statistisch leerveld is *op welk moment* leren het beste gemeten kan worden: terwijl kinderen leren (online maat) of achteraf (offline maat)? In het verleden werd alleen gebruikt gemaakt van offline maten, maar de laatste jaren is het steeds gebruikelijker om beide maten te combineren. Het gebruik van online statistisch leermaten is nog relatief nieuw. Als we bijvoorbeeld kijken naar de studies die meegenomen werden in onze meta-analyse over auditief talig statistisch leren bij mensen met en zonder TOS (Hoofdstuk 2), dan valt op dat er slechts 1 studie een online leermaat gebruikte. Dit was ook een van de redenen dat we besloten om een nieuwe kindvriendelijke online maat van auditief talig statistisch leren te ontwikkelen. In Hoofdstuk 3 laten we zien dat deze nieuwe online statistisch leermaat geschikt is om statistisch leren in kinderen tussen de 5 en 8 jaar oud te detecteren. In Hoofdstuk 4 gebruiken we deze nieuwe maat om statistisch leren te meten in kinderen met en zonder TOS.

Samenvattend concluderen we dat het statistisch leerprobleem in kinderen met TOS en de samenhang tussen statistisch leren en taalvaardigheid afhankelijk is van verschillende factoren. Dit zijn onder andere het domein (talig of niet-talig) en de modaliteit (auditief, visueel, visueelmotorisch) waarin het leren plaatsvindt, alsook de manier waarop statistisch leren gemeten wordt. We vinden een verschil in statistisch leren tussen kinderen met en zonder TOS in onze auditieve, talige taak. In het visuele niet-talige domein en het visueelmotorische niet-talige domein zijn kinderen met TOS wel gevoelig voor regelmatigheden. In de context van taalinterventies en behandeling voor kinderen met TOS kan dit betekenen dat het trainen van statistisch leervaardigheden beperkt effect kan hebben op het verbeteren van de taalvaardigheid. Het kan dus effectiever zijn om binnen de behandelcontext te focussen op het verbeteren van andere cognitieve aspecten die sterker samenhangen met taalvaardigheid dan met statistisch leren.

About the author

Imme Lammertink was born July 17, 1991, in Nijmegen, The Netherlands. She obtained a bachelor's degree (combined with a 2-years Honours Program) in Dutch Literature and Culture, and a research master's degree in Cognitive Neuroscience from Radboud University Nijmegen. Imme wrote her bachelor's thesis, under the supervision of Prof. Paula Fikkert, on sound-symbolism in Dutch toddlers. The outcomes of this study were included in a meta-analysis on sound symbolism that Imme published together with dr. Mathilde Fort, Prof. Sharon Peperkamp, Adriana Guevara-Rukoz, Prof. Paula Fikkert and dr. Sho Tsuji in 2018 (*Developmental Science*). For her master thesis Imme studied, under the supervision of Prof. Paula Fikkert, Prof. Brechtje Post, dr. Titia Benders and dr. Marisa Casillas, how Dutch and English toddlers use prosodic and lexicosyntactic cues to predict upcoming turn transitions. Imme received a *Beyond the Frontiers* grant (Radboud University Honours Academy) to conduct part of her Masters' studies at Cambridge University, United Kingdom. In 2015, Imme started her PhD project on statistical learning in children with and without Developmental Language Disorder at the University of Amsterdam. During her PhD, Imme was supervised by Prof. Judith Rispens, Prof. Paul Boersma and Prof. Frank Wijnen and she spent two months as a guest researcher with dr. Jarrad Lum at Deakin University Melbourne, Australia. At Deakin, she learned more about the use of Transcranial Magnetic Stimulation in language research and conducted a study on the ERP components sensitive to sequence learning effects on the serial reaction time task (see Lum, Lammertink, Clark, Fuelsher, Hyde, Enticott, & Ullman, 2019). In addition to publishing and presenting the results of her PhD studies in international scientific journals and conferences, Imme was also actively involved in public outreach: she regularly wrote blogs for Wetenschap.nu, participated in Science Battles, presented her work at conferences and in journals for professionals working with children with language disorders and is active as web editor and editor of the news bulletin within an association for Dutch linguistics (het WAP). Furthermore, during her PhD project Imme also obtained a University Teaching Qualification and she worked six months as junior policy officer at the Dutch Organisation for Scientific Research (in the context of the professional PhD program). In February 2020, Imme started working as a postdoctoral researcher within the Royal Dutch Kentalis Group.