Examining the contribution of statistical learning to grammar and literacy acquisition

A study of Dutch children with and without dyslexia

In this dissertation, we investigate the hypothesis that a domain-general statistical learning mechanism supports the acquisition of language, both in its spoken and in its written form. Such a statistical learning mechanism allows for the learning of abstract patterns and rules based on the statistical properties of the input (i.e. language). Our investigation includes two separable lines of research: (1) the study of the correlation between individual differences in statistical learning ability and scores on grammar and literacy, and (2) the study of group differences between Dutch-speaking children with and without dyslexia. Moreover, it applies both experimental and meta-analytical techniques.

Taken together, the results presented in this dissertation do not provide evidence for (or against) a link between a domain-general statistical learning ability and the acquisition of language and literacy skills. Therefore, it cannot be excluded that the relationship between statistical learning and language and literacy acquisition may be less strong than hypothesized. Furthermore, individuals with dyslexia likely do not have a domain-general, extensive deficit in statistical learning. More research in the form of large-scale and pre-registered studies, as well as meta-analyses, is needed in order to reach definitive conclusions regarding the contribution of (domain-general) statistical learning ability to the acquisition of language and literacy skills, both in typical and in impaired populations.

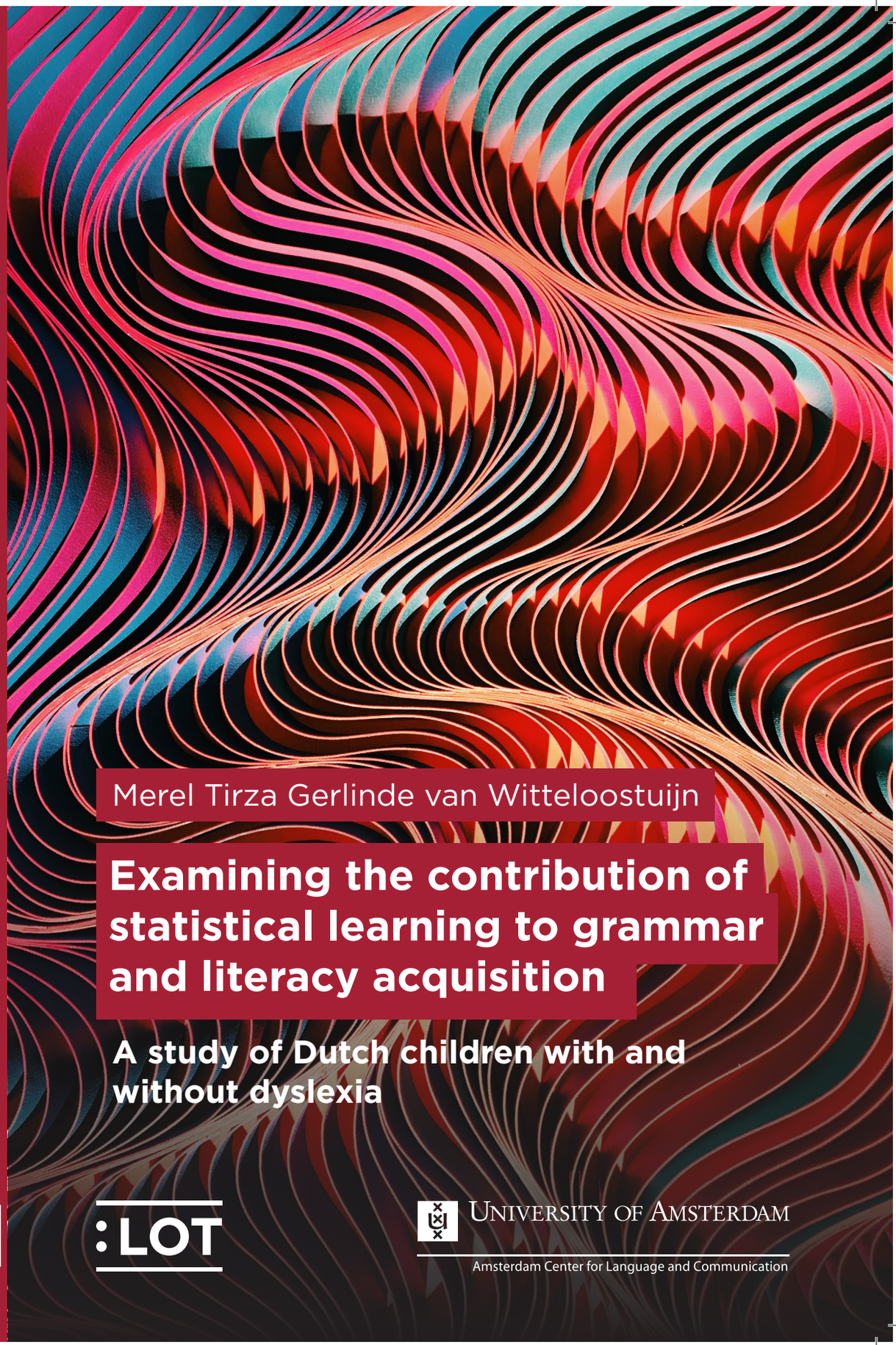
ISBN 978-94-6093-353-0

:LOT Netherlands Graduate School of Linguistics
Landelijke Onderzoekschool Taalwetenschap

:LOT 568

Merel Tirza Gerlinde van Witteloostuijn

Examining the contribution of statistical learning to grammar and literacy acquisition



Merel Tirza Gerlinde van Witteloostuijn

Examining the contribution of statistical learning to grammar and literacy acquisition

A study of Dutch children with and without dyslexia

:LOT

University of Amsterdam
Amsterdam Center for Language and Communication

Examining the contribution of statistical learning to grammar and literacy acquisition

A study of Dutch children with and without dyslexia

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: Photograph by P. Kamsirikunakorn on Unsplash

ISBN: 978-94-6093-353-0
NUR 616

Examining the contribution of statistical learning to grammar and literacy acquisition

A study of Dutch children with and without dyslexia

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 18 juni 2020, te 10.00 uur

door
Merel Tirza Gerlinde van Witteloostuijn
geboren te Maastricht

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. J.C. Schaeffer	Universiteit van Amsterdam
	Prof. dr. P.F. de Jong	Universiteit van Amsterdam
	Dr. E.H. de Bree	Universiteit van Amsterdam
	Prof. dr. R. Frost	Hebrew University of Jerusalem
	Prof. dr. P. J. Monaghan	University of Lancaster
	Dr. E. Kidd	Max Planck Institute for Psycholinguistics

Faculteit der Geesteswetenschappen

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

Table of contents

Acknowledgments

Although at times writing a PhD thesis can feel like a solitary pursuit, my experience over the last five years has felt far from lonely. Here, I would like to take the time to express my gratitude to all of the people that have, in one way or another, contributed to my PhD journey: colleagues, friends, and family.

First and foremost, I am very thankful for my group of supervisors: Paul Boersma, Judith Rispens, and Frank Wijnen. Paul, I admire your conscientious nature, your attention to detail, and your ability to explain complex statistical matters. Frank, I especially value our meetings in Utrecht, during which you gave me a fresh perspective and made me feel at home at my old university. Judith, I want to thank you for your dedication and involvement. I can't imagine how many hours I've spent talking to you in the doorway of your office; often about work-related things, but also about my (small) daily struggles, our holidays, motherhood, and my future career. I always felt welcome. Thank you to all three of you for your support and trust.

Secondly, I would like to say a big thank you to Imme Lammertink. Being part of a larger VIDI project means that I had you as my close-PhD-colleague, for which I feel really lucky. You were not only a great colleague, but I also really appreciate our times together on the train to talk about our days, our summer schools and conferences together, and your kindness and attentiveness. We're defending our theses within a week of each other and I can't wait to call you Dr. Lammertink for the first time!

Thirdly, I want to express my gratitude to the members of my reading committee: Elise de Bree, Ram Frost, Peter de Jong, Evan Kidd, Padraic Monaghan, and Jeannette Schaeffer. All of you have inspired me somewhere along the way. Elise, I have such fond memories of our time in Lancaster at the *International Dyslexia Association* conference, mostly because of the great fun we had together with Sietske. Ram, my visit to your lab in Jerusalem has brought me so much. I learned a lot from your expertise and I cherish fond memories of my time in Israel. I

enjoyed the atmosphere in the 'basement' thanks to you and your wonderful team including Noam, Louisa, and Henry.

I am, furthermore, indebted to all the wonderful student assistants that helped Imme and myself gather and/or transcribe data: Iris Broedelet-Resink, Ellen Collée, Sascha Couvee, Merel Hardeman, and Darlene Keydeniers. I am also grateful for other students, such as Sybren Spit and Veronika Vadinova, who showed an interest in our project and decided to write their theses on associated topics. A warm thank you goes out to Dirk Jan Vet; whose technical help was unmissable during the first years. Even when we came to you for the umpteenth time to adapt our experiments for piloting, you were welcoming and helpful. Likewise, this dissertation would not have been possible without all of the participating children, parents, teachers, schools and dyslexia organizations.

Within the Linguistics department of the UvA, both at the Bungehuis and the P.C. Hoofthuis, our close group of PhD colleagues has always made work feel like a home away from home. Firstly, because of our cozy office – made possible by a couch once smuggled into the Bungehuis by Iris Duinmeijer and Bibi Janssen – and its lovely inhabitants: Imme, Tessa, Tiffany, Iris, and Iris. But also because of a lot of other great colleagues on the sixth floor of the P.C. Hoofthuis, with whom we had our daily lunch and coffee(s); Marloes, Ulrika, and Marieke in particular. Being in such similar situations, we could always talk (and sometimes complain) to each other. I also want to thank our 'reading group', consisting of Hernán, Imme, Iris, Judith, Natalia and Sybren, for our many fruitful discussions. Similarly, my thanks goes out to the members of our research group *Grammar and Cognition* for the opportunity to present my project at different stages of development and for all the valuable feedback. Further, I am grateful for my mock reading committee: Desiree, Imme, Iris, Sietske, Sybren, and Sylvia.

I am very thankful to my friends for enjoying food, drinks, (escape) games, festivals, and holidays together. I would like to do so by mentioning the names of our Whatsapp groups, which to me say a lot about each individual group of friends and our shared memories. Lots and love and kisses go out to: *Embryo's voor het leven* (Anouk, Leonie, Lotte, and Suzette), *Wij ♥ Elkaar* (Bart, Chris, Gerrit, Hadassa, and Tijn), *De Eetclub 2.0* (Bertine, Cerise, Hester, and Saskia), *De*

Greppel Groep (Annemieke, Anouk, Bram, CJ, Eefje, Jolien, Marije, Marte, Martijn, Paul, Robert, Rolof, Sanne, Sofie, Thijs, Tom, and Zoé), *Biertjes? Colaatjes?* (Cerise, Ferdi, Lara, Marcel, and Marloes), and *Gastronomicon* (Jeroen and Margriet). Further, I want to mention some old colleagues, who have become friends (Chantal, Loes, and Tessel), and some new friends I've made since moving to Vianen (Angela, Ellen, Kirsten, and Ronnie). I want to highlight my oldest friend Cerise, with whom I share the best and craziest memories. My thanks also go out to my lovely paranymphs: my friend, co-linguist and co-escape-game-lover Lotte and my amazing sister (and friend) Amber.

Which leads me to my family, to whom I wish to express my gratitude in Dutch. Papa, Mama, bedankt voor jullie eindeloze liefde en steun. Waar jullie ook wonen, bij jullie voelt het als thuis. En Amber, ik prijs mezelf gelukkig met zo'n fijne zus als jij. Ik wil graag opa Klaas en opa Koos noemen, omdat ik dankbaar ben dat jullie deze mijlpaal samen met ons meemaken. Inmiddels heb ik ook een schoonfamilie bij wie ik me thuis mag voelen; bedankt Diny, Toon, Anke, Jarno, Sanne en Annemarie.

Het grootste geluk deel ik met jou, Rik. We hebben elkaar leren kennen toen ik 21 was. Nu zijn we 10 jaar verder en zijn we ouders geworden van ons mooie, lieve, grappige en eigenwijze meisje Linde. Het is moeilijk om niet in clichés te vervallen, maar met jullie een gezin vormen is het mooiste dat me is overkomen.

Author contributions

Chapter 1 – General introduction

Written by Merel van Witteloostuijn with valuable feedback from Judith Rispens, Paul Boersma, and Frank Wijnen.

Chapter 2 – An online measure of visual statistical learning

Chapter 2 is a slightly modified version of a published article: van Witteloostuijn, M.T.G., Lammertink, I.L., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (2019). Assessing visual statistical learning in early-school-aged children: The usefulness of an online reaction time measure. *Frontiers in Psychology, 10*, Article 2051.

The study was designed by Merel van Witteloostuijn in collaboration with Imme Lammertink, Paul Boersma, Frank Wijnen and Judith Rispens. The experimental design was based on previous studies by Arciuli and Simpson (2011; 2012) and by Siegelman, Bogaerts, Kronenfeld and Frost (2018). Dirk Jan Vet assisted with the technical implementation of the experiment. Merel van Witteloostuijn and Imme Lammertink recruited participants and collected data with help from Darlene Keydeniers (test assistant). Data analysis was performed by Merel van Witteloostuijn, primarily supervised by Paul Boersma. Merel van Witteloostuijn is the lead author of this manuscript, with helpful feedback from Paul Boersma, Frank Wijnen, and Judith Rispens.

Chapter 3 – Visual AGL in dyslexia: A meta-analysis

Chapter 3 is a slightly modified version of a published article: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (2017). Visual artificial grammar learning in dyslexia: A meta-analysis. *Research in Developmental Disabilities, 70*, 126–137.

The study was designed by Merel van Witteloostuijn in collaboration with Paul Boersma, Frank Wijnen and Judith Rispens. We wish to thank Jarrad Lum for meeting with us and generously providing us with tips on how to conduct a meta-analysis. Merel van Witteloostuijn performed the database searches and assessed abstracts and full-texts with valuable help from Merel Hardeman (student assistant). Data analysis was performed by Merel van Witteloostuijn, primarily supervised by Paul Boersma. Merel van Witteloostuijn is the lead author of this manuscript, with valuable feedback from Paul Boersma, Frank Wijnen, and Judith Rispens.

Chapter 4 – Statistical learning in dyslexia across three paradigms

Chapter 4 is a slightly modified version of a published article: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (2019). Statistical learning abilities of children with dyslexia across three experimental paradigms. *PLoS ONE, 14*(8), Article e0220041.

The study was designed by Merel van Witteloostuijn in collaboration with Paul Boersma, Frank Wijnen and Judith Rispens. The experiments were created in collaboration with Imme Lammertink, Paul Boersma, Frank Wijnen and Judith Rispens. Dirk Jan Vet assisted with the technical implementation of the experiments. Merel van Witteloostuijn recruited participants with dyslexia and collected their data. The control group, consisting of typically developing children, was recruited by Merel van Witteloostuijn and Imme Lammertink, and data was collected by Merel van Witteloostuijn, Imme Lammertink, and test assistants (Darlene Keydeniers, Iris Broedelet-Resink, and Sascha Couvee). Data analysis was performed by Merel van Witteloostuijn, primarily supervised by Paul Boersma. Merel van Witteloostuijn is the lead author of this manuscript, with helpful feedback from Paul Boersma, Frank Wijnen, and Judith Rispens.

Chapter 5 – The contribution of statistical learning to literacy skills

Chapter 5 is a slightly modified version of a manuscript that is currently under review: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (under review at *Dyslexia*). The contribution of individual differences in statistical learning to reading and spelling performance in children with and without dyslexia.

The study was designed by Merel van Witteloostuijn in collaboration with Paul Boersma, Frank Wijnen and Judith Rispens. Merel van Witteloostuijn recruited participants with dyslexia and collected their data. The control group, consisting of typically developing children, was recruited by Merel van Witteloostuijn and Imme Lammertink, and data was collected by Merel van Witteloostuijn, Imme Lammertink, and test assistants (Darlene Keydeniers, Iris Broedelet-Resink, and Sascha Couvee). Data analysis was performed by Merel van Witteloostuijn, primarily supervised by Paul Boersma. Merel van Witteloostuijn is the lead author of this manuscript, with helpful feedback from Paul Boersma, Frank Wijnen, and Judith Rispens.

Chapter 6 – Individual differences in statistical learning and grammar

Chapter 6 is a slightly modified version of a submitted manuscript: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (submitted to *Applied Psycholinguistics*). Grammatical difficulties in children with dyslexia: The contributions of individual differences in phonological memory and statistical learning.

The study was designed by Merel van Witteloostuijn in collaboration with Paul Boersma, Frank Wijnen and Judith Rispens. Merel van Witteloostuijn recruited participants with dyslexia and collected their data. The control group, consisting of typically developing children, was recruited by Merel van Witteloostuijn and Imme Lammertink, and data was collected by Merel van Witteloostuijn, Imme Lammertink, and test assistants (Darlene Keydeniers, Iris Broedelet-Resink, and Sascha Couvee). Spoken data collected for measures of grammar were transcribed and scored with the help of Ellen Collée (student assistant). Data analysis was performed by Merel van Witteloostuijn, primarily supervised by Paul Boersma. Merel van Witteloostuijn is the lead author of this manuscript, with helpful feedback from Paul Boersma, Frank Wijnen, and Judith Rispens.

Chapter 7 – General discussion

Written by Merel van Witteloostuijn with valuable feedback from Judith Rispens, Paul Boersma, and Frank Wijnen.

Chapter 1

General introduction

As adults, we are fascinated by children's relatively quick and seemingly effortless acquisition of their mother tongue. Before children start primary school at the age of four, they know approximately 1,500 words and are able to combine these words to produce full sentences. Naturally, their linguistic skills will continue to develop, but this early stage of language acquisition is remarkable given the fact that 4-year-old children are often still unable to complete "simple" tasks such as tying their shoe laces. One of the fundamental questions in the field of linguistics, therefore, is how children are such efficient language learners, despite the complexity of language itself and in absence of explicit instruction. Put more broadly, how children learn patterns and regularities in the world around them is a long-standing question. Central to this question is the innateness debate: are (linguistic) patterns and regularities learned purely through exposure or is such learning supported by some form of innate knowledge? The traditional nativist (or "knowledge-driven") account presupposes that innate and domain-specific knowledge is needed for language acquisition. This is usually referred to as Universal Grammar (UG; e.g. Chomsky, 1986; 1995). In contrast, input-driven accounts argue that domain-specific innate knowledge is unnecessary for language acquisition. Instead, acquisition is shaped by mere exposure to language and the employment of domain-general cognitive abilities (e.g. Tomasello, 2003). More specifically, through repeated exposure to the distributional properties of language, and a domain-general ability to (implicitly) track these distributional statistics across time and space (Frost, Armstrong, Siegelman, & Christiansen, 2015), children are thought to infer the abstract patterns and rules of their native language. This ability to learn from distributional statistics is often referred to as "statistical learning", a term first introduced by Saffran, Newport and Aslin (1996), and is argued to play an important role in language and literacy acquisition (e.g. Aslin & Newport, 2014; Romberg & Saffran, 2010; Treiman, 2018).

Although language acquisition occurs rapidly and with relative ease for most children, large individual differences in the speed and ease of acquisition

exist. At the lower end of the spectrum, between 3 to 10 percent of the general population is diagnosed with developmental language disorder (DLD) or developmental dyslexia (henceforth "dyslexia"; Leonard, 2014; Miles, 2004; Siegel, 2006). Whereas DLD is characterized by spoken (or signed) language deficits (Leonard, 2014), dyslexia is associated with deficits in the development of written language (i.e. technical reading and spelling; Snowling, 2001). In both cases, these problems occur despite normal intelligence, normal academic and social opportunities and in absence of sensory or neurological impairments (e.g. DSM-V, 2013; Snowling, 2000). Since spoken language ability and literacy skills are crucial to an individual's social and academic success, children with DLD and dyslexia are vulnerable to social and/or academic problems (e.g. Conti-Ramsden, Durkin, Toseeb, Botting, & Pickles, 2018; Humphrey & Mullins, 2002).

Over the past decades, various accounts have been put forward to explain these language-based disorders. Generally speaking, theories of DLD have focussed on problems with language processing (e.g. working memory; Archibald & Gathercole, 2006) or on specific problems in the area of grammar (Leonard, 2014), while dyslexia has often been explained through underlying problems in the area of phonology and phonological memory (e.g. de Bree, 2007; Ramus, 2003), even though non-linguistic explanations have also been put forward, such as visual problems and a specific problem in mapping letters to speech sounds (e.g. Froyen, Willems, & Blomert, 2011; Stein & Walsh, 1997). It is important to note that there is considerable overlap in the symptoms of the two disorders: many children with DLD experience problems with (technical) reading and spelling, and children with dyslexia have been shown to be delayed in spoken language development (e.g. Durkin, Fraser, & Conti-Ramsden, 2010; McArthur, Hogben, Edwards, Heath, & Mengler, 2000; Snowling & Melby-Lervåg, 2016). Moreover, comorbidity between DLD and dyslexia is high (Bernthal, Bankson, & Flipsen, 2009; Catts, Adlof, Hogan, & Weismer, 2005). These facts have led some researchers to view dyslexia and DLD as resulting from the same underlying problem, namely a domain-general learning deficit (Fawcett & Nicolson, 2019; Nicolson & Fawcett, 2007; 2011; Ullman 2004; Ullman & Pierpont, 2005; Ullman, Sayako Earle, Walenski, & Janacsek, 2019). This domain-general learning deficit is conceptualized as a problem with procedural learning, i.e. the learning of deterministic and/or probabilistic associations between adjacent or nonadjacent stimuli through repeated practice and training, which is thought to be an automatic and implicit process (Janacsek

& Nemeth, 2012; Ullman et al., 2019). Statistical learning is assumed to rely on the same brain structures that support procedural learning (i.e. the basal ganglia; Ullman et al., 2019) and has been argued to be a form of procedural learning (Qi, Sanchez Araujo, Georgan, Gabrieli, & Arciuli, 2019; Steacy et al., 2019). Since there is evidence that statistical learning plays a (perhaps critical) role in language acquisition, it is not surprising that a deficit in this type of learning is hypothesized to cause the language problems observed in children with DLD or dyslexia. An important outstanding question is how the differences between the two language-based disorders can be explained under the assumption of a common underlying statistical learning deficit (e.g. primary problems in written language in dyslexia and primary problems in spoken language in DLD), although this question will not be addressed in the present dissertation (see e.g. Bishop & Snowling, 2004).

This dissertation investigates the hypothesized relationship between statistical learning and spoken and written language acquisition using two approaches. The first of these is an individual differences approach: if statistical learning is related to language acquisition, one would expect to find correlations between individual language outcomes (e.g. grammar, technical reading, and spelling) and measures of statistical learning ability. The second approach is the comparison between impaired and unimpaired individuals: if the language difficulties observed in developmental disorders can be explained through an underlying problem with statistical learning, one would expect to find group differences between individuals with and without a diagnosis of DLD or dyslexia on statistical learning tasks. The studies presented in this dissertation are part of a project investigating both developmental disorders, but the focus here lies exclusively on children with dyslexia. The results regarding children with DLD are reported elsewhere (Lammertink, Boersma, Wijnen, & Rispens, 2019a; 2019b; 2020). Thus, the following sections discuss the role statistical learning may play in language and literacy acquisition (§1.1), and describe what is known about statistical learning in dyslexia (§1.2). Finally, §1.3 provides an overview of this dissertation's contents and an outline of its chapters.

1.1 Statistical learning and (written) language acquisition

In relation to language and literacy acquisition, statistical learning tasks often target the ability to track sequential statistics (Romberg & Saffran, 2010). These sequential statistics are typically conceptualized through transitional probabilities (TPs): a TP is the probability of event *t* given the previous event *t*-1. The ability to track TPs has been argued to play a role at many levels of natural languages: the segmentation of fluent speech into words, the detection of dependencies and other co-occurrences in sentences, and the acquisition of the language's writing system. For example, the TP between syllables that form a word is higher than the TP between syllables that cross a word boundary (e.g. in the utterance "pretty baby", the TPs from *pre* to *ty* and from *ba* to *by* are higher than the TP from *ty* to *ba*; Saffran et al., 1996). One of the first statistical learning experiment, in which infants were exposed to an artificial language in which TPs between syllables were manipulated, showed that infants are sensitive to TP structure and are able to subsequently discriminate sequences of syllables with high TPs between syllables (i.e. syllables that co-occur with a high frequency; "words") from sequences with low TPs between syllables (i.e. syllables that co-occur with a low frequency; "partwords"; Saffran et al., 1996). These findings show, firstly, that infants can track the statistical information in an artificial speech stream. Secondly, they support the possibility that infants use a statistical learning mechanism to detect word boundaries in real-life language acquisition. Following this study, statistical learning experiments have expanded to investigate the potential of this domain-general learning mechanism across multiple levels of linguistic structure. More remote relationships between linguistic elements, such as the relationship between auxiliaries and inflections on the main verb (e.g. *is* walk*ing* or *has* play*ed*), may also be supported by a general learning mechanism that allows individuals to track these more remote co-occurrences (i.e. nonadjacent dependencies; Gómez, 2002).

Most relevant to the present dissertation, which focusses on individuals with dyslexia, is the relationship between statistical learning and literacy acquisition. Orthography, besides being a stream of visual elements, is of course a representation of the sounds of spoken language (Treiman, 2018). Learning to read and spell in alphabetic languages therefore depends on the process of linking orthographic units (i.e. graphemes) with phonological units (i.e. phonemes). In

other words, literacy acquisition starts with learning which letters correspond to which sounds and vice versa (grapheme–phoneme mappings). Just as spoken language, these grapheme–phoneme mappings are known to exhibit many statistical regularities: the pronunciation of a single letter may depend on co-occurring letters. For example, the letter <c> is pronounced as a /k/ when followed by the letter <a> as in *can't*, while it is pronounced as an /s/ when it is followed by an <e> as in *cent*. Similarly, statistical patterns exist purely at the level of the orthography: some combinations of letters occur more frequently than others. For example, the doubling of consonants is more common before <ick> spellings as in *gimmick* than before <ic> spellings as in *mimic* (Samara, Singh, & Wonnacott, 2019). Although some of these grapheme–phoneme associations and spelling rules are highly regular and can be taught explicitly, others are more inconsistent and difficult to state explicitly. For this reason, implicit learning processes are thought to be involved in learning to read and spell (e.g. Sperling, Lu, & Manis, 2004; Treiman, 2018; Arciuli, 2018).

In summary, a mechanism that allows for the detection of statistical patterns (e.g. patterns of co-occurrences of syllables, verb inflection, and phoneme–grapheme correspondences) is hypothesized to facilitate the acquisition of structure in language, both in its spoken and in its written form. Although experiments usually target one level of learning (e.g. syllables or words), language learners in the real world may use this domain-general learning mechanism to track all kinds of regularities in the world around them simultaneously. Thus, statistical learning may contribute not only to detecting the frequencies and co-occurrences of speech sounds, syllables and words, but also to detecting regularities in the context of a linguistic utterance, such as physical objects in the surroundings and social cues such as a speaker's eye gaze (Romberg & Saffran, 2010).

In line with the hypothesized relationship between statistical learning on the one hand and language and literacy acquisition on the other hand, empirical studies have yielded evidence of the positive correlation between measures of statistical learning ability and measures of language and literacy skills. Performance in different statistical learning paradigms (e.g. visuo-motoric serial reaction time [SRT], visual statistical learning [VSL], artificial grammar learning [AGL], auditory statistical learning [ASL], and auditory nonadjacent dependency learning [NADL] tasks) has been shown to relate to levels of ability in various components of spoken and written language. In English-speaking adults,

individual differences in statistical learning performance have been shown to correlate not only with their comprehension of complex sentences (e.g. containing relative clauses; Misyak, Christiansen, & Tomblin, 2010; Misyak & Christiansen, 2012), but also with word reading (Arciuli & Simpson, 2012) and reading Hebrew as a second language (Frost, Siegelman, Narkiss, & Afek, 2013). Similarly, the SRT performance of English-speaking typically developing (TD) children has been found to relate to their reaction times (RTs) on a sentence–picture matching task, taken as a measure of grammatical processing (Clark & Lum, 2017), and to the proportion of passive sentences produced in a syntactic priming experiment (Kidd, 2012). In a study adopting the VSL paradigm, children's statistical learning ability was found to correlate with their comprehension of passive sentences and sentences that contain object relative clauses, as measured through accuracy on a sentence-picture matching task (Kidd & Arciuli, 2016). Regarding literacy skills, positive correlations in child participants have been reported when looking at the relationship between statistical learning and the reading of individual words (VSL paradigm: Arciuli & Simpson, 2012; SRT paradigm: Hung et al., 2019) and between ASL and sentence reading (Qi et al., 2019).

While the abovementioned results are promising, a number of studies have yielded null results regarding correlations between statistical learning tasks and language performance. In 2012, Lum and Kidd did not find a significant correlation between children's SRT performance and their accuracy on an elicitation test of the past tense. West, Vadillo, Shanks, and Hulme (2018) reported null results regarding the correlation between an SRT task and measures of literacy skills (i.e. spelling and word reading) in a large sample of English-speaking TD children. In a similar fashion, Schmalz, Moll, Mulatti, and Schulte-Körne (2019) found no evidence for (or against) a relationship between two statistical learning tasks and both word and nonword reading fluency in a sample of German-speaking adults, and Clark and Lum (2017) found no evidence for a relationship between SRT performance and word and nonword reading in children with and without DLD. This mixed pattern of findings in the field (i.e. some studies find evidence of correlations, while other studies do not) has led researchers to question the strength of the relationship between statistical learning and the performance on tasks that assess language and literacy skills (e.g. Schmalz et al., 2019). Moreover, these findings have raised doubt about our ability to assess this relationship reliably, especially in child participants (e.g.

Arnon, 2019a; 2019b; Kidd, Donnelly, & Christiansen, 2017; West et al., 2018). Methodological differences may also help explain the existence of mixed findings; if the true effect is small, it may only appear under certain methodological conditions (e.g. Elleman, Steacy, & Compton, 2019; Schmalz et al., 2019). These methodological differences may relate to the choice of statistical learning task, as different paradigms are known to target different types of structure (e.g. adjacent, nonadajcent) in different modalities (e.g. visuo-motor, visual, auditory). Individual experiments also vary on a large number of other methodological parameters (e.g. type of stimuli, length of exposure, instruction, measure of learning, etc.). Therefore, researchers have emphasized the need for studies that use a range of statistical learning tasks within a large sample of participants (Arciuli & Conway, 2018; West et al., 2018). It is also important to note that performance on statistical learning tasks is thought to be related to an individual's ability to maintain attention and to store information in (short-term, working and long-term) memory (e.g. Arciuli, 2017), and most previous studies have not considered these potential cognitive confounds (but see Qi et al., 2019; von Koss Torkildsen, Arciuli, & Wie, 2019). Furthermore, studies on the relationship between statistical learning and literacy skills have largely focussed on reading. Thereby, they have disregarded spelling, despite its theorized link with statistical learning (Treiman, 2018). The studies presented in this dissertation add to this body of research and aim to assess the relationship between statistical learning ability and performance on language and literacy skills using a range of statistical learning tasks that span different structure types and modalities while controlling for the abovementioned cognitive factors (see §1.3 for more detail).

1.2 Statistical learning and dyslexia

As mentioned previously, a deficit in the area of procedural learning has been hypothesized to be the underlying cause for dyslexia (e.g. Nicolson & Fawcett, 2007; Ullman, 2004), which has since then been extended to include problems with statistical learning (e.g. Gabay et al., 2015; Ullman et al., 2019). As statistical learning is assumed to play a critical role in the acquisition of grapheme–phoneme associations in typical development, a statistical learning deficit in dyslexia may cause less developed and less automatic grapheme–phoneme associations, which in turn may result in their difficulties in learning to read and

write. Besides providing an explanation for the observed literacy problems in dyslexia, a statistical learning deficit may account for a range of additional symptoms associated with dyslexia, including difficulties in other domains of language, such as inflectional morphology (e.g. Joanisse, Manis, Keating, & Seidenberg, 2000; Rispens & Been, 2007) and syntax (e.g. Reggiani, 2010; Robertson & Joanisse, 2010), but also in non-linguistic skills such as motor functioning (e.g. Ramus, Pidgeon, & Frith, 2003).

A number of studies report evidence of poor statistical learning abilities in individuals with dyslexia as compared to individuals without, providing support for the hypothesized statistical learning deficit in dyslexia. This holds for investigations of performance using a range of statistical learning tasks, including the SRT task (e.g. Jiménez-Fernández, Vaquero, Jiménez, & Defior, 2011; Vicari, Marotta, Menghini, Molinari, & Petrosini, 2003), ASL task (Gabay, Thiessen, & Holt, 2015), VSL task (Sigurdardottir, Danielsdottir, Gudmundsdottir, Hjartarson, Thorarinsdottir, & Kristjánsson, 2017; Singh, Walk, & Conway, 2018), and AGL task (e.g. Pavlidou & Williams, 2014). However, as previously described for correlational studies in TD participants (§1.1), mixed findings exist regarding the statistical learning performance of individuals with dyslexia. Non-significant results regarding the difference in performance between participants with and without dyslexia have been reported by studies adopting the SRT task (e.g. Kelly, Griffiths, & Frith, 2002), AGL task (e.g. Rüsseler, Gerth, & Münthe, 2006) and the NADL task (Kerkhoff, de Bree, & Wijnen, 2017). For this reason, Lum, Ullman, and Conti-Ramsden (2013) conducted a meta-analysis of 14 previously published studies that investigated the statistical learning performance of individuals with and without dyslexia, focusing on the SRT task. Their findings suggest that, on average, individuals with dyslexia are poorer learners on the SRT task when compared to age-matched individuals without dyslexia (weighted average effect size = .449, $p < .001$). Although SRT and AGL tasks have been extensively used to examine the hypothesized (domain-general) statistical learning deficit in dyslexia, less is known about other measures of statistical learning ability, such as VSL and NADL tasks. The present dissertation employs three distinct statistical learning paradigms to thoroughly test the hypothesized (domain-general) statistical learning deficit in a large sample of children with and without dyslexia.

1.3 This dissertation

The current dissertation builds on previous work and investigates the relationship between statistical learning and (spoken and written) language acquisition in children with and without dyslexia. The key hypotheses are that (1) individual differences in statistical learning ability are related to language and literacy performance, and (2) children with dyslexia perform more poorly on statistical learning tasks than their TD peers. Three experimental tasks that tap into different aspects of statistical learning ability were developed in order to test these hypotheses in a comprehensive way. Moreover, language and literacy abilities were tested at multiple levels, including not only (technical) reading and spelling but also inflectional morphology and syntax, and the studies presented in this dissertation were controlled for cognitive factors known to influence either statistical learning performance (e.g. attention, working memory) or linguistic performance (e.g. phonological memory, rapid automatized naming, vocabulary). In doing so, we aimed to examine the relationship between statistical learning on the one hand and language performance and/or dyslexia on the other hand independently of these potentially confounding factors. In the following sections, some important considerations when measuring children's statistical learning abilities are discussed (§1.3.1), followed by a presentation of the contents of this dissertation (§1.3.2).

1.3.1 Measuring statistical learning ability

Because statistical learning is assumed to be a domain-general learning mechanism, and this domain-general ability is hypothesized to play a role in the acquisition of language and literacy skills and (hence) in explaining dyslexia, the statistical learning tasks employed in the present dissertation span a range of modalities. Furthermore, they were constructed so that they target the learning of different types of statistical structures and use a combination of explicit and implicit measures of learning, in order to obtain a broad picture of children's statistical learning ability. The three statistical learning paradigms that are utilized in this dissertation (SRT, VSL, auditory NADL [A-NADL]) are described below.

The classical SRT task assesses the visuo-motoric learning of a repeated sequence (Nissen & Bullemer, 1987). A visual stimulus is repeatedly presented in one of four marked locations on a screen and without the participants' knowledge, the visual stimulus appears according to a pre-determined 10-item sequence (e.g. 4, 2, 3, 1, 2, 4, 3, 1, 4, 3, where numbers 1–4 correspond to the four marked locations on the screen). Meanwhile, participants are required to respond to the stimuli by pressing on buttons that correspond to their location on the screen. Sensitivity to this type of structured input is measured as a participant's increase in RTs when the visual stimulus no longer follows the sequence, but is instead presented in random locations. As the visual stimulus no longer appears in a fixed order, its appearance is no longer predictable, resulting in an increase in RTs. In the second paradigm, the VSL task, participants are exposed to a continuous stream of visual stimuli that, unbeknownst to them, is structured into groups of three stimuli (i.e. triplets; Saffran et al., 1996). In such a structure, the occurrence of the second and third elements of a triplet is predictable based on the first element, while the occurrence of the first element of the following triplet is unpredictable. For example, within the triplet *ABC*, *B* always follows *A* and *C* always follows *B*, but the triplet *ABC* in the stream of stimuli may be followed by a range of other triplets (e.g. *DEF*, *GHI* or *JKL*). Finally, the A-NADL task targets participants' learning of nonadjacent relationships in auditory input, while ignoring an intervening stimulus (e.g. in the string *aXb*, *a* predicts *b* and *X* is a variable intervening stimulus). In other words, whereas the adjacent relationships (i.e. from *a* to *X* and from *X* to *b*) are unpredictable, the nonadjacent relationship (from *a* to *b*) is predictable. In a typical VSL or A-NADL task, participants are tested on their sensitivity to the statistical structure subsequent to exposure. Learning is then reflected by participants' ability to distinguish test items that adhere to the statistical structure (i.e. VSL: an existing triplet such as the triplet *ABC*; A-NADL: an *aXb* item) from test items that do not. Thus, the three tasks presented here span three distinguishable modalities: the SRT task is visuo-motoric, the VSL task is visual, and the A-NADL is an auditory task. The same holds for the type of structures targeted by the three tasks: although all can be defined as statistical in nature, the SRT contains a repeatedly presented sequence of 10 items, the VSL presents four triplets in a random order, and the structure of the A-NADL is nonadjacent.

Beside spanning different modalities and statistical structure types, learning in the three statistical learning tasks is measured through a combination

of post-hoc explicit decision-making measures (i.e. "offline" measures) and measures that assess learning as it unfolds through collecting RTs to individual stimuli (i.e. "online" measures). Performance on tasks such as the VSL and A-NADL is conventionally assessed using offline measures, but the use of these measures has been questioned in recent years (e.g. Christiansen, 2019; Frost, Armstrong, & Christiansen, 2019; Kidd et al., 2017; Siegelman, Bogaerts & Frost, 2017; Siegelman & Frost, 2015). There are three main reasons to question the use of offline measures: (1) they inform researchers only about the outcome of the learning process, not about the learning process itself; (2) the learning process is assumed to be (largely) implicit, and online measures are likely to better reflect this implicit learning ability than explicit offline ones; and (3) the initial stages of the statistical learning process, i.e. the real-time encoding of the stimuli and the patterns, is not captured by offline measures (e.g. Batterink & Paller, 2017). Therefore, in all three tasks, we measured sensitivity to the statistical structure through RTs during exposure. The hypothesis when using these online measures is that participants who are sensitive to the statistical structure process predictable input faster than unpredictable input (Siegelman, Bogaerts, Kronenfeld & Frost, 2018). This idea is based on the SRT task, which is traditionally assessed through an online RT measure as explained above. To enable the collection of meaningful online RT data in the SRT and A-NADL tasks, the presentation of structured input is followed by a block of unstructured stimuli (i.e. stimuli are presented semi-randomly). Following the abovementioned hypothesis, RTs in the block of unstructured input are expected to be slower than RTs in the surrounding structured blocks (see López-Barroso, Cucurell, Rodríguez-Fornells & de Diego-Balaguer, 2016, for a similar approach to measuring the A-NADL task online with adult participants). In the VSL, RTs to individual stimuli are collected through a self-paced design (Siegelman et al., 2018), and RTs to unpredictable elements within triplets (i.e. element 1; *A* in the triplet *ABC*) are hypothesized to be processed slower than predictable elements (i.e. elements 2 and 3 within triplets; *B* and *C* in the triplet *ABC*). In addition to these online measures, the resulting knowledge about the statistical structure is assessed through offline measures in the VSL and A-NADL tasks.

To summarize, the three statistical learning tasks described here were created to provide a comprehensive view of children's statistical learning abilities by targeting different types of statistical structures in a range of modalities, and by using both on- and offline measures of learning. They were used to investigate

(1) the relationship between individual differences in learning ability and linguistic performance, and (2) potential group differences between participants with and without dyslexia.

1.3.2 Outline of dissertation chapters

This dissertation contains six further chapters. Chapter 2 is dedicated to examining the usefulness of an online RT-based measure to assess VSL performance in child participants (van Witteloostuijn, Lammertink, Boersma, Wijnen, & Rispens, 2019). Although a similar online measure has been shown to be sensitive to learning in adults (Siegelman et al., 2018), its suitability for use with child participants was previously unknown.

Chapters 3 and 4 target our second hypothesis regarding the statistical learning performance of children with dyslexia as compared to TD children. As discussed in §1.2, previous studies showed a mixed pattern of findings: some studies found significant differences in performance between participants with and without dyslexia, while others reported null results. For this reason, prior to commencing our experimental study of children with and without dyslexia, a meta-analysis was conducted (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017, see chapter 3). This meta-analysis included evidence from 13 published and unpublished studies on a measure of statistical learning (the visual AGL task) in participants with and without dyslexia. The main research question was whether the accumulated data would provide evidence for statistical learning problems in individuals with dyslexia. In similar vein, the statistical learning tasks as described under §1.3.1 are employed in chapter 4 (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019) to investigate the same research question: do children with dyslexia show poorer performance on statistical learning? If the results of chapters 3 and 4 indicate poorer performance in individuals with dyslexia on a range of statistical learning tasks, these findings will support the hypothesized (domain-general) statistical learning deficit in individuals with dyslexia (or, put more broadly, the procedural learning deficit; Nicolson & Fawcett, 2007; 2011; 2019; Ullman 2004; Ullman & Pierpont, 2005; Ullman et al., 2019).

The first hypothesis introduced in this dissertation, i.e. that individual differences in statistical learning are related to language and literacy performance, is investigated in chapters 5 and 6 (van Witteloostuijn, Boersma, Wijnen, &

Rispens, under review; van Witteloostuijn, Boersma, Wijnen, & Rispens, submitted). While chapter 5 focuses on the relationship between tasks that assess statistical learning in the visual domain (SRT and VSL) and literacy skills, chapter 6 examines the contribution of the SRT and A-NADL tasks to inflectional morphology and syntax. As explained previously, these chapters consider participant-level variables that may confound the hypothesized relationship (e.g. sustained attention, short-term and working memory, phonological processing). If a (domain-general) statistical learning ability supports the acquisition of language and literacy skills as hypothesized, we expect to find indications of this relationship in chapters 5 and 6.

The final chapter (chapter 7) recapitulates the findings of chapters 2–6 in relation to the two main hypotheses introduced here, followed by a discussion of the implications of the studies presented in this dissertation, and ending with the conclusions.

Chapter 2

An online measure of visual statistical learning*

Abstract

Purpose: Visual statistical learning (VSL) was traditionally tested through offline two-alternative forced choice (2-AFC) questions. More recently, online reaction time (RT) measures and alternative offline question types have been developed to target learning during exposure and to increase sensitivity to individual differences in adults (Siegelman et al., 2017a; 2018). We assessed the usefulness of these measures for investigating VSL in early-school-aged children. Secondarily, we examined the effect of introducing a cover task, potentially affecting attention, on children's VSL performance.

Methods: 53 children (aged 5 – 8) performed a self-paced VSL task, in which participants determine the presentation speed and RTs to each stimulus are recorded. Half of the participants performed a cover task. Subsequently, participants completed 2-AFC ("choose correct triplet") and 3-AFC ("fill blank to complete triplet") offline questions.

Results and conclusions: RTs were significantly longer for unpredictable than predictable stimuli, so we conclude that early-school-aged children are sensitive to the statistical structure during exposure, and that the RT task can measure that. We found no evidence as to whether children can perform above chance on offline 2-AFC or 3-AFC questions, or whether the cover task affects children's VSL performance. These results show the feasibility of using an online RT task when assessing VSL in early-school-aged children. This task therefore seems suitable for future studies that aim to investigate VSL across development or in clinical populations, perhaps together with behavioral tasks.

---

* This chapter is a slightly modified version of a published article: van Witteloostuijn, M.T.G., Lammertink, I.L., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (2019). Assessing visual statistical learning in early-school-aged children: The usefulness of an online reaction time measure. *Frontiers in Psychology, 10*, Article 2051.

2.1 Introduction

Research into statistical learning has shown that infants, adults, and children are able to detect statistical structure in sequences of stimuli in the world around them (e.g. Arciuli & Simpson, 2011; 2012; Fiser & Aslin, 2002; Saffran et al., 1996). Extracting statistical properties from the input is thought to be an implicit process (Perruchet & Pacton, 2006) and has been observed in both the auditory (e.g. Saffran et al., 1996) and visual modalities (e.g. Conway, Pisoni, Anaya, Karpicke, & Henning 2011; Kirkham, Slemmer, & Johnson, 2002), which has led to the suggestion that statistical learning is a domain-general learning mechanism (see Frost et al., 2015, for a review). Statistical learning has been put forward as an essential mechanism in language acquisition, which is supported by findings that have established relationships between an individual's capacity for this type of learning and his/her language and literacy proficiency (e.g. Arciuli & Simpson, 2012; Evans, Saffran, & Robe-Torres, 2009).

In the typical statistical learning paradigm, as originally employed by Saffran et al. (1996), participants are exposed to a continuous stream of visual or auditory stimuli (the *familiarization phase*). Without the participants' knowledge, the stimulus sequences are divided into triplets of co-occurring elements (e.g. the continuous string *bidakupadotigolabu* is a concatenation of three-syllable chunks/triplets *bidaku*, *padoti*, and *golabu*). The order in which these triplets occur is free. Hence, transitional probabilities (TPs) are structured such that TPs from one syllable to the next are higher for stimuli within a triplet (e.g. *daku*) than for those that span a triplet boundary (e.g. *kupa*). It is crucial that during the familiarization phase, participants are not instructed to learn or memorize the input: they either listen passively or perform a cover task that is unrelated to the statistical regularities presented to them (e.g. Arciuli & Simpson, 2011). Under these task conditions, it is assumed that the learning process is implicit.

Participants were traditionally tested on their newly acquired knowledge of the TP structure in an *offline test phase*, subsequent to the familiarization phase. Such an offline test traditionally employed two-alternative forced-choice (2-AFC) questions, in which participants are presented with one group of three-syllable stimuli that co-occurred frequently during familiarization (e.g. the probable "word" *bidaku*) and one group of three-syllable stimuli that did not co-occur frequently (e.g. the less probable "nonword" *dakupa*). Whereas for infants the

offline test phase consists of collecting listening or looking, which are used to infer a preference for either familiar (word) or unfamiliar (nonword) items, adults and children can be asked explicitly which of the two patterns of stimuli are more familiar. In the latter case, above-chance performance on the group level is taken as evidence that participants have learned the contrast between the two patterns of stimuli, taken to reflect sensitivity to the TP structure presented to them during the familiarization phase. Bertels and colleagues (2012; 2015) show that both adults and 9- to 12-year-old children who reach above-chance performance on an offline test phase had some degree of explicit knowledge of the TP structure as evidenced by confidence ratings (i.e. more confident in correct than incorrect items). Thus, although the learning process itself may be implicit, the resulting knowledge may (to some degree) be explicit.

The suitability of using offline 2-AFC questions for measuring statistical learning has been questioned, especially for use in an individual differences approach (e.g. Kidd et al., 2017; Siegelman, Bogaerts, & Frost, 2017a; Siegelman & Frost, 2015). Furthermore, Siegelman et al. (2018) argue that offline measures inform us about the *learning outcome*, but do not reveal anything about the *learning process* during the familiarization phase. Conceivably, different individuals or different populations achieve similar offline performance, but these similar performances may be the result of differing learning trajectories during familiarization (Siegelman et al., 2018). Moreover, the term "statistical learning" implies a temporal component: the assumption is that participants become increasingly responsive to the statistical structure during exposure. As explained by Batterink and Paller (2017), the initial stages of statistical learning involve the encoding of the stimuli, which gradually transforms from the encoding of individual stimuli (e.g. syllables such as *bi*, *da*, and *ku*) to the encoding of larger co-occurring units (e.g. words such as *bidaku*). This development across time indicates increased sensitivity to the structure of the sequence. Analogously, learning during familiarization will increasingly allow participants to predict upcoming stimuli, resulting in faster reaction times (RTs) to predictable stimuli as compared to unpredictable stimuli (Siegelman et al., 2018). This idea is based on the serial reaction time (SRT) task (Nissen & Bullemer, 1987), which measures participants' implicit learning of a visuo-motoric sequence as the increase in RT when participants move from structured to unstructured, and thus from predictable to unpredictable, input.

Recent studies have employed the above-mentioned ideas about online learning in novel measures of statistical learning with adult participants, providing insight into the initial and central stages of learning in adult learners, which are not tapped by offline measures (e.g. Franco, Gaillard, Cleeremans, & Destrebecqz, 2015; Gómez, Bion, & Mehler, 2011; Karuza, Farmer, Fine, Smith, & Jaeger, 2014; Misyak et al., 2010; Siegelman et al., 2018). The main aim of the present study is to extend these recent findings to child participants and to investigate the effectiveness of such an online measure with early-school-aged children, as previous studies employing online measures have focused on adult participants. Although several studies have shown children's sensitivity to statistical structure in visual stimuli (e.g. Arciuli & Simpson, 2011; 2012; Conway et al., 2011), studies combining the use of on- and offline measures during such a task are scarce (but see Qi et al., 2018). Therefore, we adopt an online RT measure of the visual statistical learning (VSL) paradigm, as developed by Siegelman and colleagues (2018), and assessed children's learning through this measure. The development of online measures is especially important for studies investigating statistical learning in early-school-aged children due to the fact that the traditional 2-AFC questions require explicit decision-making, a skill that young children have difficulties with (Bialystok, 1986). Children's performance on 2-AFC questions in VSL tasks is known to increase between the ages of 5 and 12 (Arciuli & Simpson, 2011; Raviv & Arnon, 2017; Shufaniya & Arnon, 2018). For this reason, solely using 2-AFC questions to assess early-school-aged children's performance may not provide a complete picture of their statistical learning abilities. In addition to the (implicit) online RT measure, we used two distinct (explicit) offline question types (2-AFC and 3-AFC) to investigate the usefulness of these measures with early-school-aged children. In the 3-AFC questions, participants do not choose the correct answer out of two as in traditional 2-AFC tasks, but complete a pattern by choosing the missing stimulus out of three alternatives (see e.g. Bertels et al., 2012; 2015; Siegelman et al., 2017a). Although this question type requires the participant to make an explicit judgment just as the 2-AFC questions, we hope that the 3-AFC questions are more intuitive for children and may therefore better reflect their statistical learning abilities. Before turning to our methodology and results, we will present an overview of previous studies that have adopted online measures of statistical learning.

2.1.1 Online measures of statistical learning

The most well-known task tapping statistical learning abilities through an online measure is the SRT task (Nissen & Bullemer, 1987). Whereas the SRT is informative regarding the domain of visuo-motoric sequence learning, the fine motor skills implied in this task make it less suitable for use with certain participant groups known to have less developed fine motor skills (e.g. participants with specific language impairment and/or dyslexia; Hill, 2001; Ramus et al., 2003). Moreover, learning in the SRT task likely partially reflects sensitivity to a repeated sequence of movements, rather than pure sensitivity to statistical structure in (visual) stimuli (see e.g. Robertson, 2007; West, Clayton, Shanks, & Hulme, 2019). Tasks that have been employed to investigate other types of statistical learning have largely focused on the use of offline measures of learning (e.g. VSL, artificial grammar learning, and nonadjacent dependency learning tasks). To further our understanding of the online statistical learning process, both behavioral methods such as RTs and neurophysiological methods such as Electroencephalography (EEG) have been proposed as suitable online methods of investigating the learning trajectory of these alternative statistical learning tasks. Although EEG has successfully been used to study SRT and artificial grammar learning tasks (for a review see Daltrozzo & Conway, 2014), and has recently been applied to an auditory statistical learning task similar to the one described above (Batterink & Paller, 2017), we focus here on behavioral methods employing RT-based measures of learning.

In 2010, Misyak et al. developed an online measure of statistical learning that combined exposure to an artificial grammar containing nonadjacent dependencies with features of the classic SRT task. The grammar consisted of strings of the form *aXb*, where element *a* predicts element *b* (i.e. the nonadjacent dependency) and element *X* is variable. Adult participants were exposed to an auditory speech stream that adhered to the grammar, while seeing a grid of six nonwords presented on a computer screen. Participants were required to simultaneously listen to the speech stream and click on the corresponding nonwords in the grid. Results showed that participants were faster to respond to nonwords in predictable positions (i.e. element *b* in the *a*X*b* structure) than in unpredictable positions (i.e. element *a* in the *aXb* structure). Similar to results from the SRT task, this effect of RT on position disappeared in a subsequent trial

block where the nonadjacent dependencies present in preceding blocks were violated. These results reflect participants' sensitivity to the distinction between predictable and unpredictable elements within the speech stream. Note, however, that this method is not suitable for use with early-school-aged children as it requires advanced literacy skills and is not suited for testing statistical learning in the visual modality.

Another proposed online method that uses RTs to investigate the trajectory of auditory statistical learning is asking participants to detect clicks within the speech stream whilst recording the RTs to this click detection task (Gómez et al., 2011). By presenting clicks both within and between words in the speech stream, Gómez et al. (2011) showed that participants were faster to respond to clicks between words than within words. They argued that these findings are due to participants' expectations based on the TP structure in the stream (i.e. within words TPs are high and thus participants expect the following syllables, which is not the case between words), thereby reflecting sensitivity to the TP structure.

In 2018, an online target detection task was used in two statistical learning tasks: one in the auditory and one in the visual domain (Qi et al., 2018). Participants were exposed to a stream of stimuli, which were organized into triplets (Saffran et al., 1996). In this TP structure, the occurrence of elements 2 and 3 within triplets are predictable, whereas element 1 within triplets is unpredictable (e.g. in the triplet *ABC*, elements *B* and *C* are predictable after the presentation of *A*, but the first element of the subsequent triplet, e.g. *D* in the triplet *DEF*, is unpredictable since the presentation order varies between triplets). The target task held that participants were required to respond with a button press to one out of twelve stimuli presented to them. The target was always the third stimulus in a group of three and was thus a predictable stimulus. The results showed a decrease in RTs in detecting the targets in the visual, but not auditory modality, in both adult and child (mean age = 12;2) participants, which was taken to reflect statistical learning in the visual task. In this experimental set-up, however, nothing is known about the RTs to non-target trials (i.e. the first and second stimulus in each group of three). It could be the case that a similar acceleration in RTs would appear for these stimuli, which would indicate accommodation to the task in general (i.e. a practice effect) instead of sensitivity to the TP structure during the learning process.

Finally, and most relevant to the present study, two recent studies have applied the self-paced reading method to statistical learning in the visual domain (Karuza et al., 2014; Siegelman et al., 2018). This approach allows participants to control the rate of exposure during familiarization by letting them press a button each time they want to proceed to the following stimulus. In this paradigm, RTs to each individual stimulus are recorded, allowing for the direct comparison of RTs to predictable versus unpredictable stimuli. Karuza et al. (2014) tested adult participants on a visual self-paced nonadjacent artificial grammar learning task containing strings of the form *aXb* and showed that predictable elements yielded shorter RTs than unpredictable elements, thus corroborating the findings by Misyak et al. (2010). Similarly, Siegelman et al. (2018) assessed learning in the visual triplet learning task. In line with previous findings, and following their predictions, results show that adults respond slower to unpredictable stimuli (element 1 within triplets) than predictable stimuli (elements 2 and 3 within triplets). The question of whether a similar RT measure of VSL could be employed in child research is yet unanswered.

In sum, previous studies have shown that online measures are an important tool to study learning during the familiarization phase of statistical learning experiments and provide additional insights into an individual's performance. In the present study we therefore aim to investigate whether RTs to individual stimuli during familiarization, as introduced by Karuza et al. (2014) and Siegelman et al. (2018), could be used to assess learning in early-school-aged children (perhaps in addition to traditional offline measures). There are several important differences between children and adults that should be taken into account in the assessment of their behavior, one of them being the control of attention. Young children are immature with respect to attentional control as compared to adults (Garon, Bryson, & Smith, 2008). Since attention is a critical component of statistical learning (Arciuli, 2017; Baker, Olson, & Behrmann, 2004; Toro, Sinnett, & Soto-Faraco, 2005), a secondary aim was to find out whether a cover task that attracts children's attention to the VSL task (responding to a deviating visual stimulus) influences their learning performance. Although cover tasks have been used in VSL experiments with children and adults to ensure that participants' attention is targeted to the stimulus stream (e.g. Arciuli & Simpson, 2011), the effect of the presence (or absence) of a cover task on learning performance in VSL tasks has not yet been investigated.

2.1.2 The current study

To test whether online measures of statistical learning are a useful method to investigate learning in child participants, we conducted a study of children's performance on a statistical learning task containing both online and offline measures. Our main aim was to test whether the online RT measure introduced by Siegelman et al. (2018) is able to assess statistical learning in early-school-aged children by employing a child-adapted version of their self-paced VSL task and could thus be used in addition to more traditional offline measures. The offline test phase consisted of two parts: next to the conventional 2-AFC questions, we included 3-AFC questions in which children were required to complete triplets by choosing one out of three possible stimuli (Bertels et al., 2012; 2015; Siegelman et al., 2017a). Our secondary aim was to assess the effect of a cover task on children's learning in the self-paced visual statistical learning task. Half of the participants completed the self-paced VSL task with cover task (Arciuli & Simpson, 2011), while the other half completed the same experiment without cover task. Therefore, our analyses of the self-paced VSL task were aimed to answer the following research questions:

1) Can we use the online RT measure of the self-paced VSL task to assess learning in early-school-aged children?
2) Can we use the offline test performance of the self-paced VSL task to assess learning in early-school-aged children?
3) Do children who receive a cover task during the self-paced VSL task perform differently on the on- and offline measures of learning than children who do not perform a cover task?

If early-school-aged children are sensitive to the TP structure of the stimulus sequence presented to them in the self-paced VSL task, we expect them to respond more slowly to unpredictable elements (i.e. element 1 of a triplet) than to predictable elements within triplets (i.e. elements 2 and 3), in line with the results obtained with adults (Siegelman et al., 2018). Furthermore, learning in the online measure could be reflected in an interaction between the difference in RT to unpredictable versus predictable elements and the effect of time, since learning is likely to develop during the task. Regarding the second research question, if

early-school-aged children are sensitive to the TP structure of the stimulus stream and are able to express this knowledge in an offline testing situation, we expect them to perform above chance-level on these question types (i.e. proportion correct above $^1/_2$ in 2-AFC questions and above $^1/_3$ in 3-AFC questions). As for the effect of a cover task on learning outcomes, we hypothesize that the cover task increases the attention paid to the task, thereby having a positive influence on learning. However, since Franco et al. (2015) found that paying attention to deviating stimuli in the form of a click detection task impaired (offline) performance, it could also be the case that performing the cover task is detrimental to learning.

Finally, the relationship between performance on the three measures of learning used in the present study (online RT, offline 2-AFC, and offline 3-AFC) was examined as part of our exploratory analyses. If it is the case that all measures of learning represent the same underlying construct (i.e. children's sensitivity to the TP structure), we expect to find correlations between all measures. However, we may encounter some difficulties measuring children's sensitivity to the TP structure in offline measures, as offline performance may rely on alternate processes such as explicit decision making. Therefore, this may result in the absence of a correlation between the online and offline measures. Alternatively, low correlations between on- and offline measures could be the result of differential underlying components of statistical learning (e.g. online measures may reflect implicit learning processes whereas offline measures may tap into more explicit knowledge; see e.g. Bertels et al., 2012; 2015; Siegelman, Bogaerts, Christiansen, & Frost, 2017b).

2.2 Materials and methods

2.2.1 Participants

Dutch-speaking typically developing children were recruited from grade 1 and 2 in four primary schools located in four different provinces of the Netherlands. From the original sample of 54 children, one child was excluded due to equipment failure. Thus, the final sample consisted of 53 participants (26 girls, 27 boys) aged between 5;9 and 8;7 (age in years; months, M = 7;3, SD = 0;6).

Fourteen participants attended grade 1, the remaining 29 participants were in grade 2. Twenty-five children (12 girls, 13 boys; mean age = 7;2, *SD* = 0;5) performed the VSL task without cover task, while the other 28 (14 girls, 14 boys; mean age = 7;3, *SD* = 0;7) performed the task with cover task. All participants were native speakers of Dutch, had no hearing problems, and no diagnosis of developmental dyslexia, language impairments, AD(H)D or autism according to teacher's reports. The ethical committee of the Faculty of Humanities of the University of Amsterdam approved the present study in 2016. Compliant with the regulations of the ethical committee, parents and/or legal guardians of the children attending grades 1 and 2 in the participating schools were informed about the research project through a newsletter and had the possibility to retract permission of including their child in the study up until 8 days after testing (i.e. passive consent).

2.2.2 Materials and design

The VSL task consisted of a familiarization phase and a subsequent offline test phase as is typical for statistical learning tasks. The structure of the current VSL task was similar to that used in several previous studies (e.g. Arciuli & Simpson, 2011; 2012). The task consisted of twelve visual stimuli that could be described as aliens, which were organized into four groups of three (i.e. triplets). These four triplets are referred to as *ABC*, *DEF*, *GHI*, and *JKL* (see Appendix A).

2.2.2.1 Familiarization phase

During familiarization, each alien was presented individually on the screen of a Surface 3 tablet with touch screen. Unbeknownst to the participant, each alien was part of a triplet that always occurred in the same order (i.e. in the triplet *ABC*, *B* always followed *A* and *C* always followed *B*). The four triplets were presented 24 times each, divided into four blocks comprising 6 repetitions per triplet. Four blocks were created so that children could take a short break in between blocks, which aimed to help them stay focused on the task. This resulted in a total of 96 triplets and 288 presentations of individual aliens. Two lists of randomized orders of presentation were created to control for potential effects of order of presentation. This randomization was constrained in two ways: (1) the same

triplet was not allowed to appear twice in a row (e.g. *ABC*, *ABC* was forbidden), and (2) pairs of triplets were not repeated (e.g. *ABC*, *JKL*, *ABC*, *JKL* was forbidden) (Arciuli & Simpson, 2011; Turk-Browne, Jungé, & Scholl, 2005). As a consequence, elements 2 and 3 of a triplet are fully predictable (with TP = 1 for one alien, and TP = 0 for the remaining 11, henceforth "predictable elements"), whereas element 1 of a triplet is less predictable (TP ≈ 0.4 for three aliens, TP = 0 for the remaining 9, henceforth "unpredictable elements"). Thus, TPs within triplets are high (from element 1 to element 2 and from element 2 to element 3), whereas TPs between triplets are low (from element 3 of triplet *i* to element 1 of triplet *i + 1*). Figure 2.1 illustrates the TP structure of the VSL task.

Importantly, a novel addition to the present VSL experiment was the use of an online RT measure during the familiarization phase. Following Siegelman and colleagues (2018), participants determined the speed of presentation of each individual alien themselves by pressing the space bar every time they wanted to proceed to the next stimulus. After each press on the space bar, the following stimulus appeared after 200 milliseconds. Due to time constraints during testing, presentation proceeded to the next stimulus when participants did not respond within 10 seconds and these trials were not included in the analyses. RTs for each space bar press were recorded for all participants and served as the online measure of statistical learning, which was used to investigate the effects of learning during the familiarization phase. The RTs to each individual alien were used as an online measure of learning, as it was hypothesized that, if early-school-aged children are sensitive to the TP structure, RTs to unpredictable elements (i.e. element 1 within triplets) would be slower than RTs to predictable elements (i.e. elements 2 and 3 within triplets; see Siegelman et al., 2018).

In order to investigate whether including a cover task in the VSL influenced participants' online and/or offline performance, half of the participants received a version of the VSL that included a cover task during the familiarization phase (Arciuli & Simpson, 2011; 2012). In the version of the experiment without cover task, the familiarization phase consisted of the continuous presentation of individual aliens that, unknown to the participant, adhered to the TP structure. In the version of the experiment with cover task, a deviant stimulus (the "intruder alien") was presented four times per block at random positions in between triplets (i.e. preceding 16.7% of all triplets in a block) and participants were required to press the intruder alien on the touchscreen to proceed. This intruder alien was always the same visual stimulus

that was not part of the set of 12 stimuli that were used to form the triplets. Importantly, the deviant stimulus was presented in random positions in the sequence, but only between – and thus not within – triplets. RTs to the triplet following the presentation of the intruder alien were not included in the analysis of the online measure of statistical learning, as these were likely to deviate from the overall RTs.

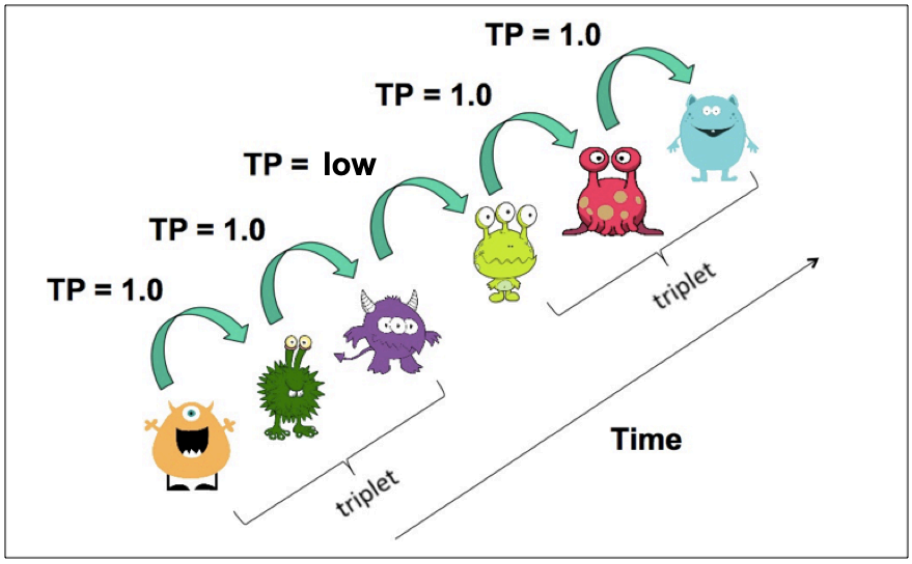


Figure 2.1. VSL task structure during familiarization. An illustration of the VSL stimuli and the triplet and TP structure.

2.2.2.2 Offline test phase

After the familiarization phase, participants were tested on their knowledge of the triplets presented to them (the "base triplets") in an offline test phase that consisted of 40 multiple-choice questions. Using the aliens of the four base triplets, four new triplets were created that had never appeared during familiarization (the "foil triplets"). These foil triplets did not violate the position of the stimuli in the base triplets (e.g. a stimulus that appeared in the first position in the base triplet, also appeared in the first position of a foil triplet) and are referred to as *AEI*, *DHL*, *GKC*, and *JBF*. Whereas the TPs between aliens within the base triplets were 1, the foil triplets were constructed from pairs of aliens that

had a TP of 0 during training. The test phase contained two parts, both containing multiple choice questions: (1) 24 2-AFC trials in which participants were asked to pick the familiar pattern ("pattern recognition" trials, chance level = $^1/_2$), and (2) 16 3-AFC trials that required the participants to complete a missing stimulus in a pattern ("pattern completion" trials, chance level = $^1/_3$). Figure 2.2 provides examples of a 2-AFC and 3-AFC test items.

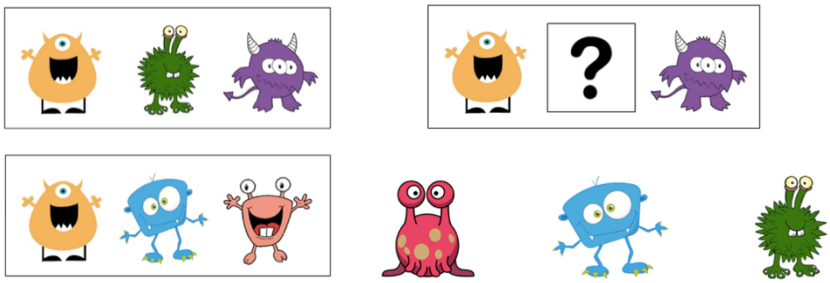


Figure 2.2. VSL offline test phase examples. Left: a 2-AFC test item. Right: a 3-AFC test item.

Test items either tested complete triplets (pattern recognition: N = 8, pattern completion: N = 8) or pairs within each triplet (pattern recognition: N = 16, pattern completion: N = 8) in order to include items that had differing properties and levels of difficulty (see Siegelman et al., 2017a). Each base triplet (e.g. *ABC*) is tested twice: in one trial it is contrasted with a foil triplet that does not contain any of the same elements (e.g. *DHL*) and in one trial with a foil triplet that contains one of the same elements (e.g. *GKC*). The same holds for each pair within base triplets (e.g. *AB* is contrasted with *DH* and *JB*). The frequency of foil triplets, pairs and single aliens was controlled for (see Appendix B for a complete overview of test items). Additionally, the position of the correct answer on the screen was controlled for and, as in the familiarization phase, two lists of randomized orders of presentation were created. Since foil triplets and pairs occurred equally frequently in the offline test phase as the base triplets and pairs, participants were not able to continue to learn during the 2-AFC questions as the opportunity to learn during testing would be equal for both base and foil triplets (Arciuli & Simpson, 2011; 2012). In all trials, possible answers were presented simultaneously on the screen and participants were instructed to choose the answer that was correct by pressing the screen. Instructions and a practice item

preceded both test phases. During the instructions and practice items, participants were encouraged to guess in case they were not certain of the correct answer.

2.2.2.3 Exit questionnaire

Following the offline test phase, half of the participants completed an exit questionnaire aimed at gaining insight into their explicit awareness of the TP structure. Consequently, information concerning explicit awareness of the VSL is available for half of the participants. The remainder of the participants completed a similar questionnaire about an auditory nonadjacent dependency learning (A-NADL) task, the results of which are described elsewhere as this task was not tested as part of the research questions of the present study (see §2.2.3 on the procedure of the present study, and see Lammertink, van Witteloostuijn, Boersma, Wijnen, and Rispens, 2019, for a discussion of the A-NADL results).

While some of the questions probed the strategies participants used, others directly asked whether participants had any explicit knowledge of the TP structure. For example, questions asked what participants were focused on during familiarization (i.e. were they focusing on the order? Or were they focused on catching the intruder in the case of receiving the version of the experiment with the cover task?), and on what strategy they applied during the test phase (e.g. did they know the answers or were they guessing?). Questions aimed at explicit knowledge of the TP structure included the question whether children noticed that the aliens stood together in groups and whether they could indicate how many aliens stood together in these groups.

2.2.3 Procedure

Each participant performed three tasks: the VSL task, a spelling test, and an A-NADL task. As mentioned, the latter tasks were not tested as part of the research questions of this article and are therefore not presented here (but see Lammertink, van Witteloostuijn et al., 2019).

The order of the tasks was controlled: half of the participants performed the VSL before the A-NADL and the other half vice versa. Additionally, half of the participants that received the VSL as their first task performed the version

with the cover task and the other half completed the version without the cover task. The same holds for those participants that received the VSL as the last task. Finally, two random orders of appearance were created to which participants were randomly assigned. The spelling task was always administered between the VSL and A-NADL tasks. In total, this resulted in a list of eight orders to which participants were randomly assigned. As mentioned in §2.2.2.3, once participants had completed all tasks, they were asked several questions probing their explicit awareness of the structure of the last task (VSL or A-NADL) they performed.

Prior to the familiarization phase of the self-paced VSL, participants were informed that they would see aliens standing in line one at a time and that they were waiting to go home in a space ship. They were instructed to send each alien home by pressing the space bar and were informed that the next alien standing in line would appear automatically. Importantly, they were told that some of the aliens really like each other and would stand in line together. Participants were instructed to watch each alien closely and to pay attention to the order of the aliens, because they would receive questions about this later (these instructions were in line with those provided in studies by Siegelman and colleagues, personal communication). Following these instructions, participants would practice the task during a practice phase containing 12 randomly ordered aliens in order to familiarize them with the procedure. The aliens included in the practice phase were different stimuli than those used in the familiarization phase. In the version of the VSL with cover task, participants received additional instructions regarding the intruder alien. The intruder alien was depicted on the screen and participants were told that this was an intruder alien that was not allowed on the spaceship. When participants saw this intruder alien, they would have to scare it away by touching it on the screen. This was followed by an additional practice round of 12 randomly ordered aliens and 3 randomly placed intruder aliens, during which participants were instructed to pay attention to the order of the aliens and to scare away the intruder aliens. Before completing the offline test phase, children were reminded of the fact that some aliens liked each other and stood in line together and were told they would receive some questions about this. An overview of the original Dutch instructions, with English glosses, is given in Appendix C.

The VSL task lasted approximately 10 minutes in total, depending on individual reaction times to the aliens in the familiarization phase and the subsequent multiple-choice questions. In between blocks of the familiarization

phase, participants had a break in which they could choose a sticker for their diploma. In the version of the task with the cover task, feedback was given on the number of times the participant caught the intruder alien. The exit questionnaire lasted approximately 3 minutes.

Children were individually tested in a quiet room at their school in a test session that lasted approximately 60 minutes. Each participant received stickers on a diploma as a reward for their participation. The VSL task was programmed and ran using E-prime 2.0 software (Psychology Software Tools, 2012; Schneider, Eschman, & Zuccolotto, 2012) on a Surface 3 tablet with touchscreen and keyboard. Instructions were recorded by a female native speaker of Dutch and played over headphones (Sennheiser HD 201).

2.2.4 Scoring and analysis

For more detail on our on- and offline analyses and the model outcomes, you can access the raw data, *R* Markdown and/or HTML files through the following link to our project page on the Open Science Framework (OSF): https://osf.io/ej32s/.

2.2.4.1 Online reaction time data

Prior to analysis, unreliable measurements were removed from the raw RT data. As mentioned, RTs to the triplet following the appearance of the intruder alien in the cover task were removed, as these reaction times are likely to deviate from the other responses (16.7% of the data for children who performed the task with detection cover task). For similar reasons, responses to the first triplet of each of the four blocks of the experiment were excluded from analysis (4.2% of data). Finally, responses faster than 50 milliseconds were removed from the dataset as these reflect cases in which the participant pressed the space bar without processing the stimulus (2.1% of data; element 1: $N = 89$, element 2: $N = 106$, element 3: $N = 86$).

Following pre-processing of raw RTs, the online RT data were analyzed using linear mixed effect models by applying the *lme4* package (Version 1.1-13; Bates, Maechler, Bolker, & Walker, 2015) for *R* software (R Core Team, 2019).

The dependent variable was the RT to each individual alien and was fitted as a function of the within-participant predictors Element (element 1, element 2, and element 3 within triplets) and Time (repetitions 1 – 24 of the triplets, which was centered and scaled), and Cover (yes or no cover task) as the between-participants predictor. Since the age of children varied between 5;9 and 8;7, Age (centered and scaled) was entered as an exploratory between-participants predictor. The two random orders of the task were also entered into the model to take away any variance associated with this contrast (Random Order 1 and Random Order 2). The model contained the maximal random effect structure that did not result in (near-)perfect correlations between the random effects (see Barr, Levy, Scheepers, & Tily, 2013) and contained by-subject and by-item[1] random intercepts and by-subject random slopes for Element and Time[2] and by-item random slopes for Cover. Age was not entered as by-subject random slopes, since this predictor naturally correlates perfectly with the by-subject intercepts. Note that the *lme4* package provides *t*-values for linear mixed effect models. Confidence intervals (CIs) and the associated *p*-values were calculated through the "profile" function (*lme4* package) and a "get.p.value" formula created for this purpose (see OSF).

2.2.4.2 Offline accuracy data

Responses on the offline test phase were coded as 1 (correct) or 0 (incorrect) for both the 2-AFC pattern recognition questions (maximum score = 24 correct) and the 3-AFC pattern completion questions (maximum score = 16 correct). Results are presented as the proportion of questions answered correctly, ranging from 0 to 1, such that chance level for the 2-AFC questions is $1/2$ and for the 3-AFC questions is $1/3$. None of the responses in the offline test phase were removed from analysis.

---

[1] Item in the online model refers to the 12 individual aliens used in the experiment.

[2] The by-subject random slopes for the interaction between Element and Time were removed from the model, as these random slopes correlated perfectly with the by-subject intercepts indicating that the model was overparameterized. Removing these random slopes was licensed, since the interaction between Element and Time was not significant. Removal did not decrease the fit of the model ($\chi^2 = 1.333$, $df = 11$, $p = .9998$) or change its main outcomes.

Offline accuracy data were analyzed using generalized linear mixed effects models for the 2-AFC and 3-AFC questions separately. The dependent variable was the accuracy of each test item (coded as 1 or 0) and was fitted as a function of Cover (yes or no cover task), Random Order (1 or 2) and Age (centered and scaled) as the between-participants predictors. The models contained by-subject random intercepts. The effect of cover task or age is interpreted as significant if the CI of the log odds does not contain zero.

2.2.4.3 Relationship between on- and offline measures

In order to investigate the relationships between the three measures used in the present study, we ran exploratory correlational analyses using the "cor.test" function with Pearson method in *R*. For the online RT measure, an individual measure of learning was calculated for each participant such that response times to predictable elements were subtracted from RTs to unpredictable elements (RT Element 1 – [RT Element 2 + RT Element 3 / 2]; see Siegelman et al., 2018). Positive individual RT difference scores thus indicate sensitivity to the TP structure, as these indicate faster responses to predictable than to unpredictable elements. For the offline measures, raw accuracy scores on the 2-AFC and 3-AFC questions were used in correlational analyses.

2.3 Results

We will first focus on the online RT measure in §2.3.1, followed by the results of the offline accuracy in §2.3.2. These sections will present confirmatory results, which answer our research questions, and subsequently address several exploratory results obtained through our linear mixed-effects analysis. Additional exploratory analyses, i.e. investigations of correlations between the different measures, are presented in §2.3.3. The exploratory results describe either unexpected findings or findings for which no prior hypotheses were constructed (cf. Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012). The results regarding the exit questionnaire are of a purely descriptive nature and are presented in §2.3.4.

Importantly, as we used multiple measures in assessing our research questions, all CIs aimed at answering our research questions were Bonferroni-corrected for multiple testing. Thus, CIs were separately adjusted for effects pertaining to evidence of online learning (research question 1), offline learning (research question 2), and the effect of the presence or absence of the cover task (research question 3). To keep the overall false detection rate at 0.05, statistical significance for confirmatory effects regarding research question 1 was determined using 97.5% CIs (i.e. the CI corresponding to a false detection rate of $0.05/2 = 0.025$), since two outcomes could provide evidence regarding online learning (i.e. the difference in RTs between predictable and unpredictable elements and this difference in RTs in interaction with Time). Similarly, 97.5% CIs were used for research question 2, since two distinct offline measures were used in the present study (2-AFC and 3-AFC questions). Finally, significance regarding research question 3 was determined using 98.75% CIs (i.e. the CI corresponding to a false detection rate of $0.05/4 = 0.0125$), since all four measures could provide an answer regarding the effect of a cover task on learning. For exploratory results we report 95% CIs.

Supplementary analyses were run including the order of the tasks (VSL or A-NADL first) as a predictor in our models, as requested by an anonymous reviewer (see OSF for files containing the supplementary analyses). Task order was found not to interact with the on- and offline measures of learning (all t and z values < 1.8). Therefore, we collapse the results from the two testing orders in our presentation of the results.

2.3.1 Online reaction time data

2.3.1.1 Online reaction time data: confirmatory results

In order to answer the first research question of whether children are sensitive to the TP structure present during familiarization, we ran the linear mixed effect model as explained in §2.2.4.1. The effect that is crucial to answering this research question is whether participants responded differently to unpredictable elements (Element 1) than predictable elements (Element 2 and 3) within triplets. Thus, the three levels of the within-participant predictor Element were coded into orthogonal contrasts such that the first contrast ("Element 1 vs. Elements 2 and

3", with Element 1 coded as - $^2/_3$ and Elements 2 and 3 coded as + $^1/_3$) estimated how much the RTs to predictable element 1 within triplets across the task differ from the mean RTs to unpredictable elements 2 and 3, which will allow us to answer our research question. The second contrast of the predictor Element estimated how much the RTs to element 2 differed from the RTs to element 3 (i.e. the two unpredictable elements, with Element 2 coded as - $^1/_2$ and Element 3 coded as + $^1/_2$), the results of which are described under §2.3.1.2 explaining our exploratory findings. The secondary effect that could answer our first research question is the interaction between the difference in RT to predictable versus unpredictable elements and Time (i.e. repetitions of triplets in the experiment), as an increase in the difference between predictable and unpredictable elements by time would indicate increasing responsivity to the TP structure across the experiment. Our second research question regarding the effect of the cover task was tested through interactions between the effect of the orthogonally contrast-coded predictor Cover (with no cover coded as - $^1/_2$ and cover coded as + $^1/_2$) and the abovementioned effects of learning (i.e. the two-way interaction between Cover and the contrast "Element 1 vs. Elements 2 and 3" or the three-way interaction with the contrasts "Element 1 vs. Elements 2 and 3" and Time).

The model was first run on raw RTs, but the resulting model's residuals were non-normally distributed. Thus, we attempted using log-transformed RTs and normalized RTs to improve the data's suitability for analysis using linear mixed effects models. Normalization was performed by sorting all N observations in increasing order, then replacing each observation by the $(r - 0.5) / N$ quantile of the normal distribution, where r is the ranking number of the observation; we consequently obtain values that can be interpreted as optimally distributed z-values. Through inspections of Quantile–Quantile ("QQ") plots of the model's residuals, it was decided that normalized RT data resulted in the best approximation of normally distributed residuals (for more detail: see the *R* markdown and/or HTML file containing all analyses on the OSF). For this reason, analyses were run on normalized RT data and the model estimates are expressed as changes in z-values (Δz) from one level of the predictor to the next.

Figure 2.3 presents the normalized RTs to elements 1, 2 and 3 within triplets over the four blocks of the experiment. Note that the normalized RTs in Figure 2.3 are averaged over blocks, which deviates from the way the analysis was conducted (i.e. on normalized RTs and using a continuous Time predictor

as explained in §2.2.4.1). As hypothesized, analysis of normalized RTs reveals that RTs to the unpredictable element 1 within triplets are significantly longer than the mean RT to both predictable elements 2 and 3 ($\Delta\tilde{\chi}$ = -0.058, *SE* = 0.022, *t* = -2.605, 97.5% CI = [-0.114 … -0.002], *p* = .021), reflecting that early-school-aged children are sensitive to the TP structure presented in the VSL task. The model estimate of the interaction with Time was not significantly different from zero ($\Delta\tilde{\chi}$ = -0.004, *SE* = 0.011, *t* = -0.328, 97.5% CI = [-0.028 … +0.021], *p* = .74). An overview of all model estimates is presented in Table 2.1. The same model was run on raw and log-RT data, resulting in similar *t*-values for the effect of unpredictable element 1 versus both predictable elements 2 and 3 (*t* = -2.074 and *t* = -2.590 respectively). Thus, the reported effect of the predictability of elements within triplets on RTs is stable across models. We did not find evidence for the effect of Element changing over the time course of the task. Figure 2.4 and Figure 2.5 provide more information regarding the time course of the experiment: Figure 2.4 plots the normalized RTs for unpredictable (Element 1) and predictable (Element 2 and 3) stimuli across repetitions of triplets (1–24), while Figure 2.5 plots the online measure of learning (i.e. difference score: normalized RT Element 1 – mean normalized RT Element 2 and 3) across repetitions of triplets (based on Figure 3 in Siegelman et al., 2018; p. 702).

Our secondary research question pertains to the effect of cover task: do early-school-aged children who receive the self-paced VSL task with a cover task respond differently from children who perform the task without a cover task? Whether the version of the task made a difference in participants' sensitivity to the TP structure is reflected in the interaction between the between-subjects predictor Cover and the first Element contrast ("Element 1 vs. Elements 2 and 3"). This interaction model estimate did not significantly differ from zero ($\Delta\tilde{\chi}$ = 0.021, *SE* = 0.023, *t* = +0.940, 98.75% CI = [-0.036 … +0.079], *p* = .35). Equally, the three-way-interaction with Time also did not differ significantly from zero ($\Delta\tilde{\chi}$ = -0.003, *SE* = 0.022, *t* = -0.156, 98.75% CI = [-0.057 … +0.051], *p* = .88). We therefore have no evidence that early-school-aged children perform the online RT task with a cover task differently than the version without a cover task.

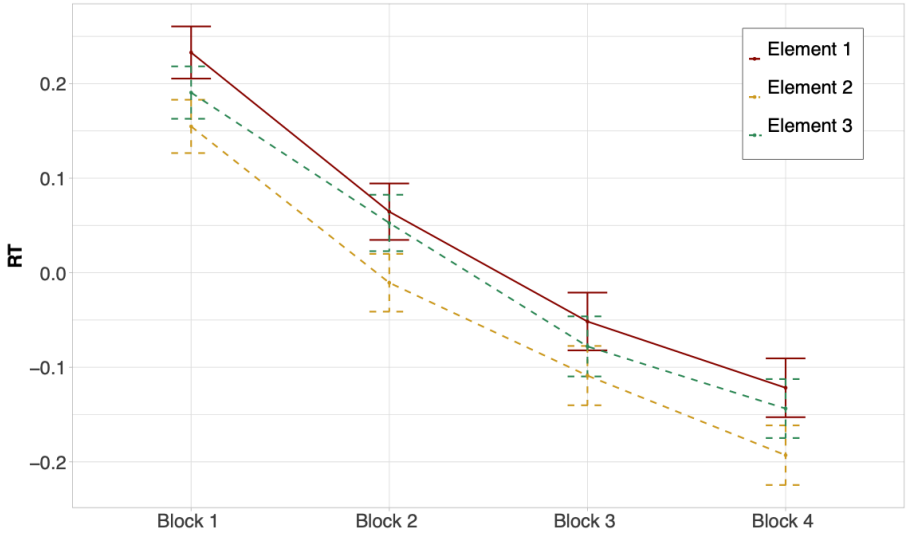


Figure 2.3. Descriptive results of the online RT data per block. Mean normalized RT (+/- 1 *SE*) to element 1 (unpredictable) and to elements 2 and 3 (predictable elements) are plotted per block of the experiment.

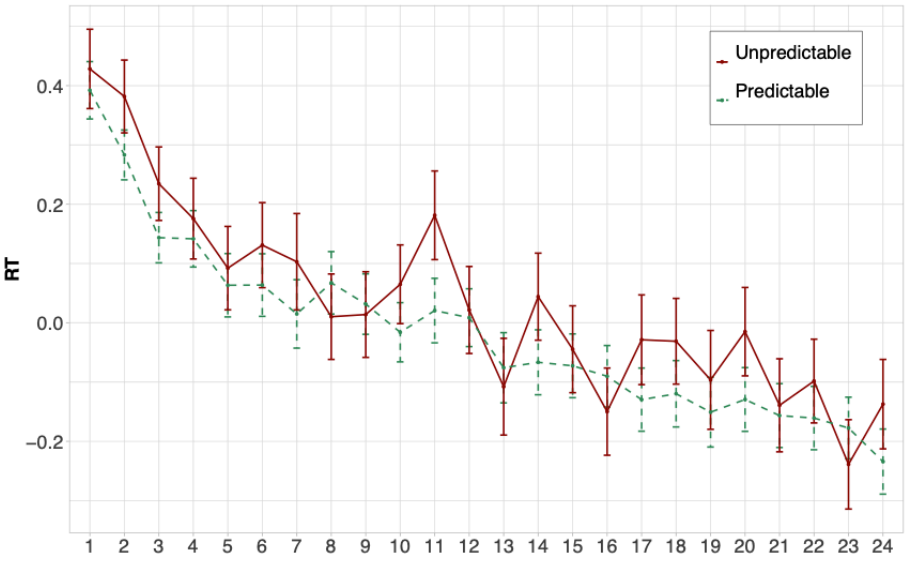


Figure 2.4. Descriptive results of the online RT data per repetition. Mean normalized RT (+/- 1 *SE*) to unpredictable elements (element 1) and predictable elements (average of elements 2 and 3) are plotted per repetition of triplets during the experiment.

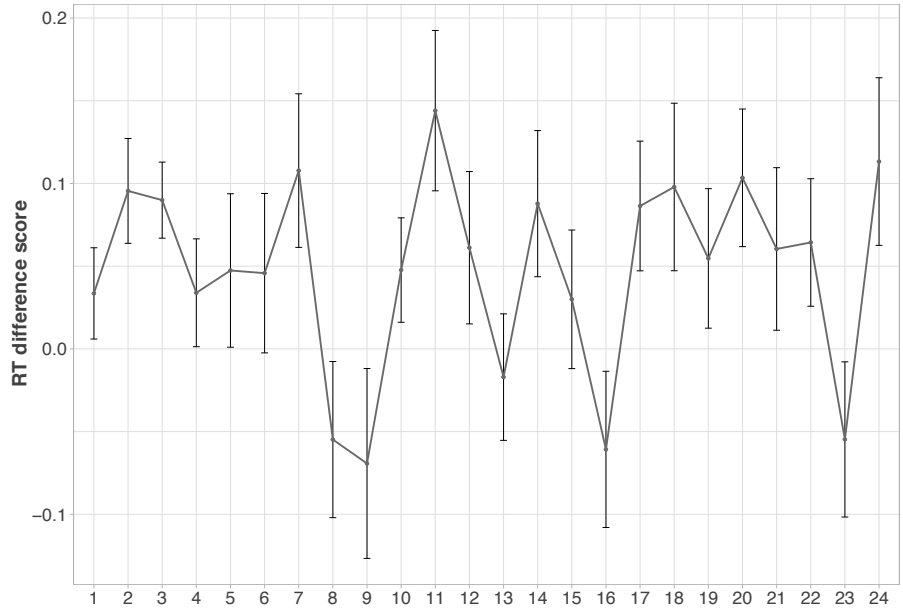


Figure 2.5. Descriptive results of the online RT data as a difference score. Mean normalized RT to unpredictable elements (element 1) minus mean normalized RT to predictable elements (average of elements 2 and 3) plotted per repetition of triplets during the experiment.

2.3.1.2 Online reaction time data: exploratory results

Besides allowing us to answer our research questions, the RT model provides some interesting exploratory results that are also evident in the normalized RTs presented in Figures 2.3, 2.4 and 2.5. Firstly, related to the TP structure of the task, we found that RTs to predictable element 2 within triplets were shorter than RTs to predictable element 3 within triplets, an effect that almost reaches significance ($\Delta\tilde{x}$ = 0.053, *SE* = 0.026, *t* = +2.056, 95% CI = [-0.002 … +0.108], *p* = .058). If this effect were real, this would mean that the difference between elements 1 and 2 is greater than the difference between elements 1 and 3, which may tell us that children predict element 2 more easily than element 3, although both element 2 and element 3 have a TP of 1 (see §2.2.2.1). As requested by an anonymous reviewer, an additional figure was created plotting the time course of

the experiment as in Figure 2.5 but excluding element 3 (i.e. normalized RT element 1 – normalized RT element 2; see supplementary materials on our OSF project page and see Appendix D).

Table 2.1. Fixed effects of the online normalized RT model, reporting on 13004 observations by 53 participants across 12 items (i.e. aliens).

	Estimate ($\Delta\tilde{z}$)	**Standard Error (SE)**	***t*-value**
(Intercept)	-0.002	0.098	-0.019
El1 vs. El2 and 3*	**-0.058**	**0.022**	**-2.605**
El2 vs. El3	*+0.053*	*0.026*	*+2.056*
*Time**	*-0.146*	*0.028*	*-5.184*
Cover	-0.142	0.195	*-0.726*
Age	*+0.269*	*0.105*	*+2.563*
El1 vs. El2 and 3 : Time	**-0.003**	**0.011**	**-0.328**
El2 vs. El3 : Time	+0.002	0.013	**+**0.125
El1 vs. El2 and 3 : Cover	**+0.021**	**0.023**	**+0.940**
El2 vs. El3 : Cover	-0.014	0.027	-0.508
El1 vs. El2 and 3 : Age	*-0.012*	*0.012*	*-0.949*
El2 vs. El3 : Age	+0.010	0.014	+0.718
El1 vs. El2 and 3 : Time : Cover	**-0.003**	**0.022**	**-0.156**
El2 vs. El3 : Time : Cover	+0.027	0.025	+1.090
El1 vs. El2 and 3 : Time : Age	*+0.004*	*0.012*	*+0.375*
El2 vs. El3 : Time : Age	+0.009	0.013	+0.688

Note. Model estimates that differ significantly from zero are indicated with an asterisk (*); those that differ marginally significantly from zero are indicated with a cross (†). El = Element. Estimates that were used to answer the research question are marked in bold; those explained under exploratory results are marked in italics.

Secondly, we see that RTs overall, thus ignoring effects of TP structure, significantly decrease as a function of Time ($\Delta\tilde{z}$ = -0.146, *SE* = 0.028, *t* = -5.184, 95% CI = [-0.203 … -0.090], $p = 3.11 \cdot 10^{-06}$). This effect of time on RTs is to be expected, as participants respond faster overall as a result of them adapting to the task and needing less time to process each individual stimulus. Finally, regarding the exploratory between-participants predictor Age (ranging between 5;9 and 8;7), the model shows that older children had significantly slower RTs overall ($\Delta\tilde{z}$ = 0.269, *SE* = 0.105, *t* = 2.563, 95% CI = [+0.131 … +0.481], *p* =

.0052), likely due to the fact that the older children in our sample have more developed academic skills and are therefore better at focusing on the task at hand. More importantly, however, we find no significant interactions between participants' age and the difference in RTs to predictable versus unpredictable stimuli or a three-way interaction between age, predictability and time ($\Delta\tilde{\chi}$ = -0.012, *SE* = 0.029, *t* = -0.949, 95% CI = [-0.036 … +0.012], *p* = .34, and $\Delta\tilde{\chi}$ = .004, *SE* = 0.012, *t* = 0.375, 95% CI = [-0.018 … +0.027], *p* = .71 respectively).

2.3.2 Offline accuracy data

2.3.2.1 Offline accuracy data: confirmatory results

Following the familiarization phase, participants performed an offline test phase consisting both of pattern recognition (2-AFC, *N* = 24) trials and pattern completion (3-AFC, *N* = 16) trials. Descriptive statistics show that participants scored between .250 and .750 correct on 2-AFC trials (*M* = .514, *SD* = .11) and between .060 and .880 correct on subsequent 3-AFC trials (*M* = .381, *SD* = .18). Figure 2.6 shows the descriptive individual and group results on the offline accuracy data for both question types.

The generalized linear mixed effects models were run on the accuracy data as explained in §2.2.4.2. The first research question was whether children can learn the TP structure presented in the VSL task, as measured by their accuracy on the offline test phase. In order to answer this question, we examined whether participants' accuracy exceeded chance level (i.e. exceeded $^1/_2$ on 2-AFC and/or $^1/_3$ on 3-AFC questions). The 2-AFC and 3-AFC model estimated that participants scored .015 and .037 above chance level respectively (2-AFC: probability intercept = .516, 3-AFC: probability intercept = .376). In both cases, this performance was found to not differ significantly from chance, as the correctness probability CIs included the task's chance probabilities (2-AFC: 97.5% CI = [+.480 … +.551], *p* = .31, and 3-AFC: 97.5% CI = [+.319 … +.429], *p* = .095). Hence, we find no evidence of above-chance performance in early-school-aged children on either 2-AFC or 3-AFC questions.

Related to our secondary research question regarding the effect of the cover task, no significant effect of cover task was found on either of the offline measures (2-AFC: odds ratio estimate = 0.927, 98.75% odds CI = [0.673 …

1.274], p = .54, and 3-AFC: odds = 1.140, 98.75% CI = [0.673 … 1.945], p = .52). Similar to our findings in the online RT measure, we cannot conclude that early-school-aged children perform the self-paced VSL with a cover task differently than the task without a cover task.

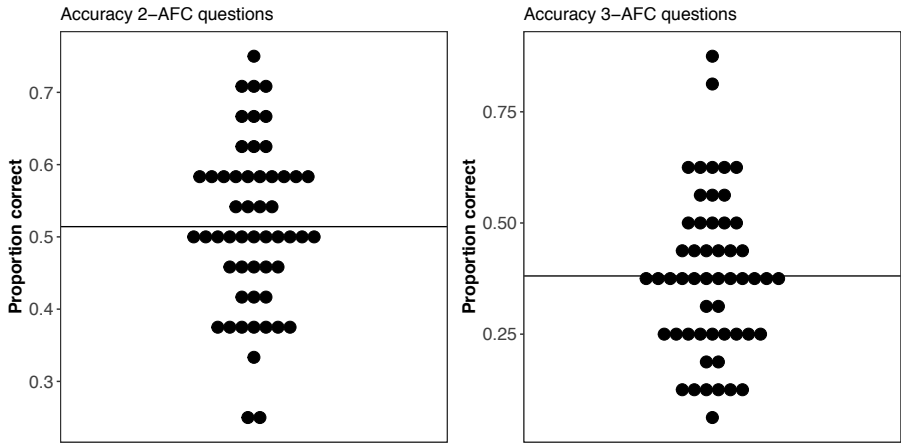


Figure 2.6. Descriptive results of the offline accuracy data. Left: distribution of scores for the 2-AFC questions (chance level = $^1/_2$). Right: distribution of scores for the 3-AFC (chance level = $^1/_3$) questions. Dots indicate individual mean accuracy scores; black lines represent overall group means.

2.3.2.2 Offline accuracy data: exploratory results

The offline models provide us with exploratory findings regarding the effect of age on performance. No significant effect of age was found on either of the offline measures (2-AFC: odds ratio estimate = +0.097, 95% CI = [-0.037 … +0.233], p = .15, and 3-AFC: odds ratio estimate = +0.131, 95% CI = [-0.103 … +0.373], p = .27). Again, in line with our findings in the online model, we find no evidence that age influences the performance of early-school-aged (between 5;9 and 8;7 years of age) children's performance on the self-paced VSL used in the present study.

2.3.3 Relationship between on- and offline measures

As mentioned in §2.2.4.3, we explored the relationship between on- and offline measures used in the present study. Since we found no effects of time on our online RT measure, the individual RT measure of learning was calculated using the normalized RTs to all stimuli presented during the experiment (normalized RT Element 1 – [normalized RT Element 2 + normalized RT Element 3 / 2]).

The results show that the two offline accuracy measures correlate significantly with one another (r = .274, t[51] = 2.031, p = .048), and neither of the offline accuracy measures correlate significantly with the online RT measure (2-AFC: r = .188, t[51] = 1.367, p = .178, and 3-AFC: r = .157, t[51] = 1.139, p = .26).

2.3.4 Exit questionnaire

Subsequent to the offline test phase, half of the participants received a short exit questionnaire (N = 24, mean age = 7;4). During familiarization, most children reported paying attention to the aliens' features (e.g. the color or the number of eyes, N = 11) or to the intruder when performing the VSL with cover task (N = 6). Five children did not give a clear answer, while the final two claimed to have paid attention to the order in which the aliens appeared. When asked whether children noticed that the aliens continuously appeared in the same groups, the majority of participants responded "no" (N = 14), whereas five participants said they did notice the order but could not explain any of the groups when shown pictures of the aliens. Only one participant could recall a single correct triplet and the four remaining children recalled incorrect (or foil) triplets. Most children said they had to guess the answers (N = 11) during the offline test phase, while others reported having memorized the correct answers (N = 4), or "just knowing" them (N = 7). The remaining two children were unable to answer this question. Finally, a large number of children thought groups of aliens consisted of either two or three aliens (N = 11), which reflects the use of both pairs and triplet items in the offline test phase. The other thirteen children either reported all groups consisted of two (N = 3), three (N = 6) or four (N = 1) aliens, two to four aliens (N = 1) or had no idea (N = 2). To summarize, the exit questionnaire did not provide

evidence of explicit strategies during familiarization or of explicit, verbalizable knowledge of the TP structure as a result of the experiment as a whole.

2.4 Discussion

In the present study, we aimed to test whether a self-paced VSL task using an online RT measure (in addition to traditional offline questions) is a useful method to investigate statistical learning in early-school-aged children. Previous work by Siegelman and colleagues (2018) has shown the suitability of such a measure for adults, but no study to date has replicated their findings with child participants. In accordance with our hypothesis, results revealed that children between 5;9 and 8;7 years old were sensitive to the TP structure during familiarization as reflected by slower RTs to unpredictable (element 1) versus predictable elements (elements 2 and 3) within triplets. We did not find evidence of an influence of the time course of the experiment on this sensitivity to predictable versus unpredictable stimuli. The reported effect of predictability is in line with previous studies with adult participants showing faster responses to predictable than unpredictable elements in statistical learning tasks, argued to reflect a difference in processing speed between predictable and unpredictable stimuli (Karuza et al., 2014; Misyak et al., 2010; Siegelman et al., 2018). The lack of an interaction with time is supported by other studies reporting that learning takes place early on during exposure (e.g. Hedenius et al., 2013). Similarly, in their investigation of the self-paced VSL with adults, Siegelman et al. (2018) report significant learning as early as after 7 repetitions of triplets. Importantly, this study demonstrates that early-school-aged children show similar sensitivity to predictability during exposure to an statistical learning task. Additionally, the online measure provides information that goes beyond the traditional offline 2-AFC (and 3-AFC) questions, for which we did not find evidence of above chance-level performance. So, while the offline accuracy data do not provide conclusive evidence for sensitivity to TP structure in early-school-aged children, the online RT measure does. This finding highlights the importance of using online measures (possibly in addition to offline measures) when investigating statistical learning in children. Moreover, the fact that the online RT measure of the self-paced VSL task has now been shown to be sensitive to children's learning abilities allows future studies to compare performance across development using the same task.

The data presented here could not determine whether 5- to 8-year-old children exceed chance level on the 2-AFC questions. We cannot reject the possibility that the failure of the 2-AFC (and 3-AFC) task could simply be due to chance (the design does not make it possible to directly compare the sensitivities of the three tasks). However, the failure could also be due to low sensitivity of the task when used with young children, which leads to difficulties in reliably measuring learning using the 2-AFC task in this population. Since the CI of the learning effect on the 2-AFC task ranged from .480 to .551, and the upper bound is thus only a performance of .551, we can cautiously conclude that if a learning effect on 2-AFC questions exists in early-school-aged children, it is a very small effect. Additionally, we found no improvement with age in this younger age group. These difficulties with assessing the VSL abilities of young children through the 2-AFC task have been reported before in the literature. In studies that employ a similar VSL task structure as presented here, significantly above-chance learning has been reported in children (Arciuli & Simpson, 2011; 2012). However, whereas children in Arciuli and Simpson (2011) were aged between 5;6 and 12;6 (M = 9;5), and between 5;10 and 12;5 (M = 9;1) in their 2012 study, children in our study were tested within the lower spectrum of their age ranges (i.e. between 5;9 and 8;7, M = 7;3). In their investigations of the effect of participant- and task-related variables on learning performance in a multiple linear regression analysis, Arciuli and Simpson (2011) found that VSL abilities develop between ages 5 and 12: learning performance on the 2-AFC task increased with age. These findings have been replicated in two other samples of children between 5 and 12 years of age, revealing higher mean performance on 2-AFC questions of a VSL task as a function of age (Raviv & Arnon, 2017; Shufaniya & Arnon, 2018). Although these findings may be interpreted as development of VSL *abilities* in these age groups, they may in fact reflect the difficulties of *measuring* children's abilities using offline measures (or, alternatively stated, they may reflect the development of the ability to make judgments involved in offline measures). This is what our results suggest, since we find evidence of sensitivity to the VSL structure in our online RT measure but no evidence of learning in our offline measures. Our results therefore underline the difficulties in using offline questions with early-school-aged children and underline the importance of using different measures in children, especially in younger age groups, to tap into their sensitivity to structure in statistical learning tasks. Early-school-aged children, as opposed to adults (and infants), may be

more likely to develop incorrect strategies when answering offline questions (e.g. focusing on the visual features of the stimuli, as we saw from the exit questionnaire) and are likely more susceptible to distractions during a complex task such as answering 2- and 3-AFC questions. Future research investigating (the development of) VSL in children could apply the online RT measure of learning as proposed here (in addition to offline measures) to obtain a more complete picture of children's statistical learning abilities.

The secondary aim of this study was to assess the effect of a cover task on children's performance in the self-paced VSL. Although we hypothesized that the inclusion of a cover task should attract children's attention to the task, thereby enhancing performance, we did not find any evidence of a positive effect of including a cover task on the offline or online performance of children. Additionally, whereas Franco et al.'s (2015) study reported that paying attention to a deviating stimulus during familiarization impaired adult participants' offline performance, we do not find evidence for a detrimental effect of our cover task on children's VSL performance either. Based on our findings, we cannot conclude whether early-school-aged children are affected by the presence or absence of the cover task in a VSL task as the one reported on here. Note that, although the cover task was designed to ensure children's attention to the VSL task (see also Arciuli & Simpson, 2011), it may be the case that it did not affect children's attention overall and therefore no evidence of an effect on VSL performance was found. Future studies that aim to investigate the potential effect of a cover task on VSL performance should include an independent measure of attention payed to the task overall to control for this possibility.

Finally, we explored the relationships between the on- and offline measures of learning used in the present study, revealing a relationship between children's performance on the two distinct offline question types as expected. We found no evidence of a relationship between the online RT measure of learning and offline performance on either 2-AFC or 3-AFC questions. This lack of correlation between online and offline statistical learning measures has been reported before (e.g. Franco et al., 2015; Misyak et al., 2010) and has several possible explanations. Firstly, although both online and offline measures are assumed to measure statistical learning in general, they may tap into different stages or different aspects of the learning process. Whereas online measures assess participants' (implicit) sensitivity to the TP structure as it is presented to them, offline measures evaluate participants' ability to make explicit judgments

about stimuli subsequent to exposure (e.g. Franco et al. 2015; Siegelman et al. 2018). Therefore, performance on these two separable processes may not be related to one another. As mentioned by Misyak et al. (2010), the online measure is a more implicit and indirect measure of learning, while the offline measure is more explicit and direct. The two types of measures may therefore be "functionally dissociable" (Cohen et al., 1990; Destrebecqz & Cleeremans, 2001; Willingham, Nissen, & Bullemer, 1989). This lack of correlation makes even more sense in the current context of early-school-aged children, since young children are known to have difficulties with explicit decision making (Bialystok, 1986). This may have resulted in the lack of evidence of above-chance performance observed in the present study, which in turn may hinder the investigation of the relationship between the different measures of learning in the self-paced VSL task. Offline measures that are more sensitive to the learning outcome of young children need to be developed in order to further explore these relationships in child participants. For example, more indirect and implicit offline measures as developed by Bertels et al. (2012; 2015) may be suitable for future research with early-school-aged children.

Although the current results regarding the online measure of learning in the self-paced VSL are very promising, we see some room for improvement. Importantly, the observed effect of predictability on children's response times was small and the difference in response times to predictable and unpredictable stimuli varied greatly between individuals. Moreover, we found no evidence of learning developing over time (i.e. an interaction between the measure of learning and the time course of the experiment, expressed as repetitions of triplets). Such an effect of time on learning would be expected theoretically, since it is assumed that participants become increasingly sensitive to the statistical structure as exposure enfolds (e.g. Batterink & Paller, 2017; Siegelman et al., 2018). While the online RT measure appears suitable for group analyses as presented in the current study, the methodology may need to be improved on in order to apply it in an individual differences approach or to investigate the time course of learning in more detail. As suggested by Siegelman et al. (2018), the presented behavioral methods may be used in combination with neurobiological methods such as EEG in order to gain more insight into the online learning process of individuals. Furthermore, methodological changes to the current design may improve the sensitivity of measuring learning online and may allow for closer inspections of the time course of learning. For example, the lack of an interaction between

learning and time in the present study may be the result of the introduction of blocks in the experiment or of participants' lack of attention to the task towards the end. While these blocks were introduced in order to keep children's attention and motivation to the task, they may have hampered the measurement of the online time course of learning by interrupting the continuous learning process. Additionally, children might need further encouragement to continuously pay attention to the stream of stimuli in this type of statistical learning tasks.

Recently, attention has been paid to the nature of the learning mechanisms underlying performance on statistical learning tasks (e.g. Siegelman, Bogaerts, Armstrong, & Frost, 2019). Learning in tasks such as the VSL presented here could be the result of sensitivity to local TPs (i.e. between pairs of stimuli) or may alternatively follow from sensitivity to more global TP patterns (i.e. "chunks" or triplets; see Siegelman et al., 2019, for a discussion). In their study of adult participants, Siegelman et al. (2019) show that participants apply both types of learning, and the reliance on one or the other differs across participants. As can be gleaned from Figure 2.3 and the *p*-value of 0.058 reported in §2.3.1.1, the results from the present study may suggest a larger difference between element 1 as compared to element 2 than as compared to element 3 within triplets, which may be indicative of larger sensitivity to local than to global TPs (i.e. pairs versus triplets) in child participants. Please note that this is highly speculative, since the present study was not set up to differentiate between these two learning mechanisms. However, this line of research opens up avenues for further investigations of the interplay between differing learning mechanisms, both in adult and in child participants. Moreover, the online RT measure of learning is a tool that is potentially useful in such explorations (see also Siegelman et al., 2019).

In sum, the present study underlines the importance of developing novel sensitive measures of statistical learning appropriate for child research and looking beyond traditional offline questions when investigating statistical learning in (early-school-aged) children. Online measures cannot only reveal sensitivity to statistical regularities during familiarization that offline questions cannot, but also have the potential to inform us about the learning trajectories of participants in different statistical learning tasks, although further research is needed to reach this goal. The RT measure of learning presented here provides an implicit, online measure that can detect sensitivity to TP structure during exposure. The self-paced VSL has thus been shown to be a useful tool in

assessing learning in children and could be further developed and adapted for future studies investigating developmental patterns of VSL or for use in clinical populations (perhaps besides more traditional offline measures). For example, a number of studies have shown impairments in the area of statistical learning in individuals with developmental language disorder (DLD) and dyslexia (see e.g. Evans et al., 2009; Gabay, Thiessen, & Holt, 2015). Online measures could provide further information regarding the differences in performance between such populations and their neurotypical peers. Future research could investigate the use of the self-paced VSL for an individual differences approach by exploring the relationship between the online sensitivity to TP structure of individual participants and their performance on language measures.

Chapter 3

Visual AGL in dyslexia: A meta-analysis*

Abstract

Purpose: Literacy impairments in dyslexia have been hypothesized to be (partly) due to an implicit learning deficit. However, studies of implicit[3] visual artificial grammar learning (AGL) have often yielded null results. The aim of this study is to weigh the evidence collected thus far by performing a meta-analysis of studies on implicit visual AGL in dyslexia.

Methods: Thirteen studies were selected through a systematic literature search, representing data from 255 participants with dyslexia and 292 control participants (mean age range: 8.5 to 36.8 years old).

Results and conclusions: If the 13 selected studies constitute a random sample, individuals with dyslexia perform worse on average than non-dyslexic individuals (average weighted effect size = 0.46, 95% CI [0.14 … 0.77], $p = .008$), with a larger effect in children than in adults ($p = .041$; average weighted effect sizes 0.71 [sig.] versus 0.16 [non-sig.]). However, the presence of a publication bias indicates the existence of missing studies that may well null the effect. While the studies under investigation demonstrate that implicit visual AGL is impaired in dyslexia (more so in children than in adults, if in adults at all), the detected publication bias suggests that the effect might in fact be zero.

---

* This chapter is a slightly modified version of a published article: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (2017). Visual artificial grammar learning in dyslexia: A meta-analysis. *Research in Developmental Disabilities, 70*, 126–137.

3 The learning process targeted by the visual AGL is referred to as "statistical learning" throughout the rest of this dissertation (see e.g. Frost et al., 2019, for motivation why the concept of statistical learning stretches to include AGL paradigms). However, the original publication of chapter 3 used the term "implicit learning" and this was left unchanged in this chapter.

3.1 Introduction

Individuals with dyslexia have severe and persistent difficulties with learning to read and spell. These difficulties occur despite normal intelligence, adequate educational and socio-economic opportunities, and in absence of sensory or neurological impairment (DSM-IV, 2000).[4] A generally accepted hypothesis is that the persistent difficulties with written language result from a core deficit in phonological processing and, specifically, phonological awareness (see Melby-Lervåg, Lyster, & Hulme, 2012 for a meta-analysis). Phonological awareness is the ability to detect and manipulate phonological segments of words (Shankweiler et al., 1995) and is related to the ability to map letters to sounds, which in turn affects the ability to learn to read and spell. Individuals with dyslexia also experience difficulties in other areas of language. Subtle problems have been reported in the area of inflectional morphology (e.g. pluralization and tense marking: Joanisse et al., 2000; subject-verb agreement: Rispens & Been, 2007; Rispens, Roeleven, & Koster, 2004) and syntax (relative clauses: Mann, Shankweiler, & Smith, 1984; Stein, Cairns, & Zurif, 1984, passive sentences: Stein, Cairns, & Zurif, 1984; binding: Waltzman & Cairns, 2000). Additionally, dyslexia is associated with a range of non-linguistic cognitive dysfunctions, including impairments in visual and auditory processing (Stein & Walsh, 1997; Tallal, 2004), attention (Facoetti, Paganoni, & Lorusso, 2000), motor functioning (Ramus et al., 2003), and verbal working memory (Gathercole et al., 2006; Gathercole & Baddeley, 1990; Swanson & Jerman, 2007).

Several theories have attempted to define the underlying deficit that accounts for the range of problems experienced by individuals with dyslexia. One recent approach is explaining dyslexia as the result of a problem with implicit learning (see Nicolson & Fawcett, 2007; Ullman & Pierpont, 2005). The term implicit learning refers to the process through which humans extract rules and regularities from visual and auditory sequences available in the environment. Importantly, this happens in absence of awareness.

---

[4] Note that in the fifth version of the DSM (DSM-V, 2013), dyslexia is included under the umbrella term "Specific Learning Disorder".

3.1.1 Implicit learning and literacy acquisition

Many studies have related implicit learning abilities to different aspects of language acquisition: the ability to segment words from continuous speech (Saffran et al., 1996), the acquisition of phonological categories and phonotactics (Nicolson & Fawcett, 2007; Wijnen, 2013), vocabulary acquisition (Evans et al., 2009; Yu, 2008), and more general language processing (e.g. passives: Kidd, 2012; relative clauses: Misyak et al., 2010). Most important to the present discussion is the relationship between implicit learning and the acquisition of literacy skills, as these are the skills most affected in individuals with dyslexia. Learning to read and spell involves the mapping between letters and sounds (grapheme-to-phoneme mapping), which requires phonological awareness and knowledge of the orthographic system. This mapping, and the writing system in general, comprises many regularities. For example, a single letter (e.g. <c>) can map onto several phonemes (e.g. /k/, /s/). Whether the letter <c> is realized as a /k/ or an /s/, depends on co-occurring letters (e.g. the letter <c> followed by the letter <a> generally results in the realization of the phoneme /k/ as in *can't*, but in the phoneme /s/ when followed by an <e> as in *cent*). In other words, the writing system consists of a "set of correlations that determine the possible co-occurrences of letter sequences, which eventually result in establishing orthographic representations" (Frost et al., 2013, p. 2). Although some of these regularities in written language are taught explicitly, it seems plausible that children's literacy acquisition is aided by implicit learning through exposure to written language.

Previous research has suggested a link between implicit learning and literacy skills in the typically developing (TD) population (e.g. Apfelbaum, Hazeltine, & McMurray, 2013; Arciuli & Cupples, 2006; Arciuli & Simpson, 2012; Frost et al., 2013; Pacton, Fayol, & Perruchet, 2005; Spencer, Kaschak, Jones, & Lonigan, 2014). For example, TD children apply orthographic regularities in pseudo-word spelling (e.g. in French, /ɛt/ is more often written as <ette> after *–v* than after *–f*), which reflects their implicit knowledge of single letters and letter combinations (Pacton et al., 2005). Similarly, Pacton and colleagues (2001) show that French-speaking TD children are sensitive to the orthographic constraints of the positions of double consonants (e.g. *xevvu* is more acceptable than *xxevu*). Additionally, correlational studies have established a link between performance

on implicit learning tasks and reading in English (Arciuli and Simpson, 2012), reading in Hebrew as a second language (Frost et al., 2013), and a variety of literacy-related skills including oral language, vocabulary and phonological processing (Spencer et al., 2014). Using a linear regression analysis, Ise, Arnoldi, Bartling, and Schulte-Körne (2012) showed that children's performance on a visual artificial grammar learning (AGL) task, a measure of implicit learning which will be explained in more detail below, predicts their performance on a spelling task. Together, the abovementioned studies suggest there is a relationship between implicit learning and (the acquisition of) literacy skills in typical populations.

3.1.2 Implicit learning in dyslexia

A number of studies have investigated the hypothesis that individuals with dyslexia have problems with implicit learning, which affect their literacy skills. Several tasks have been deployed to investigate implicit learning skills in dyslexia. Examples include the serial reaction time (SRT) task (e.g. Deroost, Zeischka, Coomans, Bouazza, Depessemier, & Soetens, 2010; Menghini et al., 2010; Vicari et al., 2005), the alternating SRT task (Hedenius et al., 2013), as well as visual AGL tasks (e.g. Ise et al., 2012; Pothos & Kirk, 2004; Rüsseler et al., 2006). Although both the SRT and AGL paradigm are methods used to investigate implicit learning, the type of structure learned in each paradigm differs (greatly). Whereas the SRT measures a motoric response to visual sequences and is stimulus-bound (i.e. no generalization rule can be abstracted from the sequence), the visual AGL measures rule learning from visual input. While numerous studies report implicit learning difficulties in individuals with dyslexia (e.g. Du & Kelly, 2013; Ise et al., 2012; Jiménez-Fernández et al., 2011; Vicari et al., 2005), others do not find evidence for such a deficit (e.g. Deroost et al., 2010; Menghini et al., 2010; Pothos & Kirk, 2004; Rüsseler et al., 2006).

Because of these mixed results, Lum et al. (2013) performed a meta-analysis on 14 studies that investigated implicit learning in individuals with dyslexia using the SRT paradigm. Their results show that implicit sequence learning, as measured by the SRT task, is significantly poorer in people with dyslexia than in non-dyslexic controls (average weighted effect size 0.45, $p < .001$). Thus, these results indicate a deficit in implicit visuo-motor learning in

dyslexia. In the current study we investigate whether individuals with developmental dyslexia are also affected in visual artificial grammar learning. If individuals with dyslexia have difficulties with implicit learning across the board, group differences should be found using both the SRT and AGL paradigms. However, it could also be the case that poor performance by individuals with dyslexia on the SRT task is due to a specific *motor* learning deficiency, as dyslexia has previously been associated with motor problems (e.g. Fawcett & Nicolson, 1995; Ramus, 2003; Ramus et al., 2003). In that case, one would not necessarily also expect difficulties in the area of visual AGL learning.

3.1.3 Visual AGL in dyslexia

Visual AGL refers to an experimental design that investigates participants' ability to implicitly learn rules from mere exposure to sequences of visual stimuli generated by these rules. First introduced by Reber (1967), the visual AGL paradigm involves structured sequences that can be presented as letters or abstract shapes. In visual AGL tasks, sequences are generated on the basis of a (finite state) grammar that determines which stimuli can and cannot succeed one another (Figure 3.1). In the example depicted in Figure 3.1, from the node S_2, the sequence can proceed either to S_4 (a triangle) or S_5 (a diamond), but not back to S_1.

The AGL task typically consists of two phases: a training and a test phase. In the training phase, participants are exposed to a set of structured sequences. Importantly, in the implicit version of the AGL task that is explored in the current meta-analysis, participants are not informed about the presence of the structural rules in the input. The exposure during the training phase can be either passive (i.e. participants are merely exposed to stimuli) or active (i.e. participants are instructed to memorize strings of stimuli and repeat them afterwards).

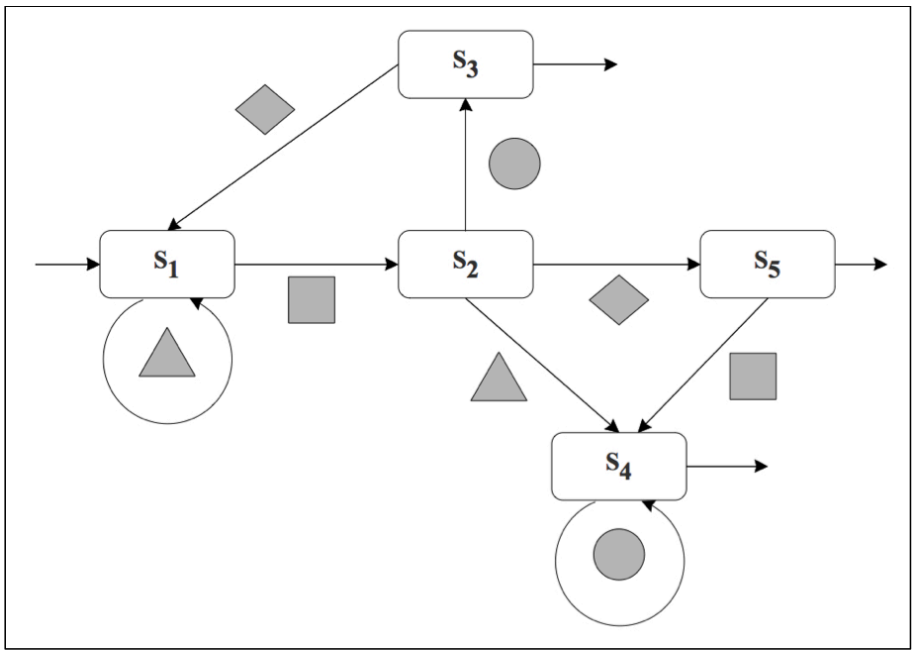


Figure 3.1. An illustration of a grammar used in AGL experiments (after Laasonen et al., 2014; original grammar by Abrams & Reber, 1988) to generate grammatical sequences (e.g. triangle, square, triangle, circle or square, diamond, square) and ungrammatical sequences (e.g. triangle, diamond, circle, diamond or square, triangle, square).

At the beginning of the test phase, participants are often informed that certain rules guided the presentation of stimuli during the training phase. Subsequently, they are tested on their ability to distinguish sequences that adhere to the artificial grammar (grammatical strings) from sequences that do not (ungrammatical strings). Typically, recognition of grammatical strings is tested within a grammaticality judgment task in which participants are requested to specify whether single strings are grammatical or ungrammatical. Other studies adopt a two-alternative forced choice paradigm, where participants are presented with two strings, one grammatical and one ungrammatical, and have to indicate which of the two strings belongs to the grammar. Performance above chance level (50%) during the test phase is taken as evidence that participants have learned the rules of the underlying grammar.

Several studies using the visual AGL paradigm have reported learning deficits in dyslexia among adults (Laasonen et al., 2014; Kahta & Schiff, 2016) or

children (Ise et al., 2012; Pavlidou, Williams, & Kelly, 2009; Pavlidou & Williams, 2014). In each of these studies, this deficit is reflected by significantly lower accuracy scores in the group of individuals with dyslexia as compared to a control group. Several other studies failed to show a significant effect of dyslexia in children (Nigro, Jiménez-Fernández, Simpson, & Defior, 2016) or adults (Pothos & Kirk, 2004; Rüsseler et al., 2006). These differences in degrees of significance might be due to chance (i.e. sampling error), because no direct statistical comparisons were ever made between the studies. However, differences in group effects might also reflect genuine differences between the studies. Here we will speculate on several factors that may help explain such genuine differences between individual studies.

Firstly, the age of the participants may influence the results of individual studies, as several studies have reported that implicit learning improved with age in typical populations (e.g. Arciuli & Simpson, 2011; Maybery, Taylor, & O'Brien-Malone, 1995, but see Jost, Conway, Purdy, & Hendricks, 2011). In a meta-analysis investigating SRT performance, Lum and colleagues (2013) found smaller differences between participants with and without dyslexia for studies involving adult as opposed to child participants when certain sequences of stimuli were used (second-order sequences) or when the exposure phase was longer. However, no previous studies have examined the developmental trajectory of AGL in individuals with dyslexia.

Secondly, the use of either linguistic or non-linguistic stimuli may influence the difficulty of the task, especially for participants with dyslexia. Linguistic stimuli include visually presented letters (e.g. Ise et al., 2012; Nigro et al., 2016; Rüsseler et al., 2006), whereas non-linguistic experiments have used abstract shapes (e.g. Laasonen et al., 2014; Nigro et al., 2016; Pavlidou et al., 2009; Pothos & Kirk, 2004). The results are mixed: several studies report impaired learning within a AGL task involving linguistic stimuli (i.e. letters, e.g. Ise et al., 2012; Samara, 2013), while others do not find evidence for an effect of dyslexia (e.g. Nigro et al., 2016; Rüsseler et al., 2006). Similarly, studies have yielded mixed results in AGL tasks with non-linguistic stimuli such as abstract shapes (evidence for learning deficits: Laasonen et al., 2014; Pavlidou, Kelly, & Williams, 2010; Pavlidou & Williams, 2014, no evidence for learning deficits: Nigro et al., 2016; Pothos & Kirk, 2004).

Thirdly, the training method potentially affects participants' performance. As mentioned, the training phase generally includes one of two

possible methods: passive exposure (Du, 2013; Laasonen et al., 2014; Nigro et al., 2016) or active memorization (e.g. Ise et al., 2012; Rüsseler et al., 2006; Samara, 2013). Active training may lead to better learning, as participants are more focused on the stimuli. Whether the observed differences in results between the studies are genuine or due to chance is one of the questions that the present paper tries to address.

Thus, mixed results exist for the visual AGL paradigm: whereas several studies report significant differences between participants with and without dyslexia (e.g. Laasonen et al., 2014; Ise et al., 2012), others do not (Rüsseler et al., 2006; Nigro et al., 2016). Schmalz, Altoè, and Mulatti (2017) conducted a meta-analysis on a subset of studies investigating visual AGL in dyslexia. They report significantly poorer performance by participants with dyslexia (average weighted effect size 0.47). However, at the same time they are careful in their interpretation and state "[…] publication bias and questionable research practices result in an inflated effect size" (p. 9). As no meta-regression analysis was performed, the authors could not quantitatively explain the differences in effect size between studies.

The primary aim of the present meta-analysis is to extend the findings by Schmalz and colleagues (2017) to a larger set of (unpublished) studies and determine whether the accumulated evidence indicates a difference in performance on visual AGL between individuals with and without dyslexia. By doing a systematic literature search and by including a number of unpublished studies, we want to provide a more complete update on the strength of the evidence regarding the association between dyslexia and a deficiency in visual artificial grammar learning. Additionally, we aim to investigate the effect of certain methodological variables through a meta-regression analysis. These variables include the age of participants and the nature and complexity of the task used, which potentially help explain heterogeneity in results of individual studies. Factors included in the analysis are (a) age (adult or child participants), (b) stimulus type (letters or abstract shapes), and (c) type of training method (passive exposure or active memorization).

3.2 Method

3.2.1 Literature search

We identified studies published up until September 2016 through searches in PubMed, PsycInfo, ERIC, MEDLINE, CINAHL, and LLBA databases. Additionally, the OATD database was searched for unpublished work in the form of theses and dissertations. A complete overview of keywords used for each of the databases can be found in Appendix E. In addition to database searches, references of included articles were reviewed. Finally, the CogDevSoc and LinguistList mailing lists were used to inquire whether subscribers knew of unpublished data (deadline response: September 2016).

3.2.2 Study selection

Figure 3.2 depicts the selection of studies according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines in a flow diagram (Moher, Liberati, Tetzlaff, & Altman, 2009). Out of all 229 records found, 143 duplicates were removed. Subsequently, one researcher examined the abstracts of 86 unique studies. Studies had to fulfill several selection criteria for inclusion in the present meta-analysis. First, only studies that had administered a visual AGL task were considered. The main reason for a focus on visual AGL studies is to eliminate modality as a possible cause of heterogeneity in results. Second, the experiment had to address implicit learning, i.e. participants were not to be informed of the presence of rules in the input. Third, studies had to include two groups of participants, one group of individuals with dyslexia and one group of non-dyslexic controls.

Fifty-six records were removed after screening the title and abstract because they did not meet the abovementioned selection criteria. An additional 19 records were removed from the sample on the basis of full-article screening, thus leaving eleven records for inclusion in the present review and meta-analysis. Two out of the eleven records (Ise et al., 2012; Nigro et al., 2016) involved two experiments with distinct participant groups that were included separately in the present meta-analysis, resulting in 13 individual effect size calculations. For the

remainder of the present meta-analysis, we will refer to the number of individual effect size calculations as the number of studies included (13). A second researcher performed identical database searches and assessed all abstracts and full texts. For 28 out of 30 full-text studies the reviewers independently came to the same conclusions regarding inclusion in the present meta-analysis (high inter-rater reliability: Cohen's kappa = .851). Consensus on the remaining two records was reached through discussion of the contents.

Note that articles did not have to have been published in peer-reviewed journals in order to be included in our meta-analysis. This means that conference papers or posters, unpublished results and dissertations could be included in the final sample (under the category "other" in Figure 3.2). This was done to minimize the possibility of a publication bias. 10 out of 12 records in this category were found through the OATD database (*Open Access Theses and Dissertations*), of which 2 are included in the final sample (Samara, 2013; Du, 2013). The other two were discovered through personal communication with authors or were presented at the *Interdisciplinary Advances in Statistical Learning* conference (2015, San Sebastian).[5] At the time of analysis, two out of thirteen individual effect sizes included in the present meta-analysis were unpublished.

[5] This resulted in the inclusion of a poster presentation by Kahta and Schiff, which was later published in 2016 and was thus later also found through the systematic literature search.

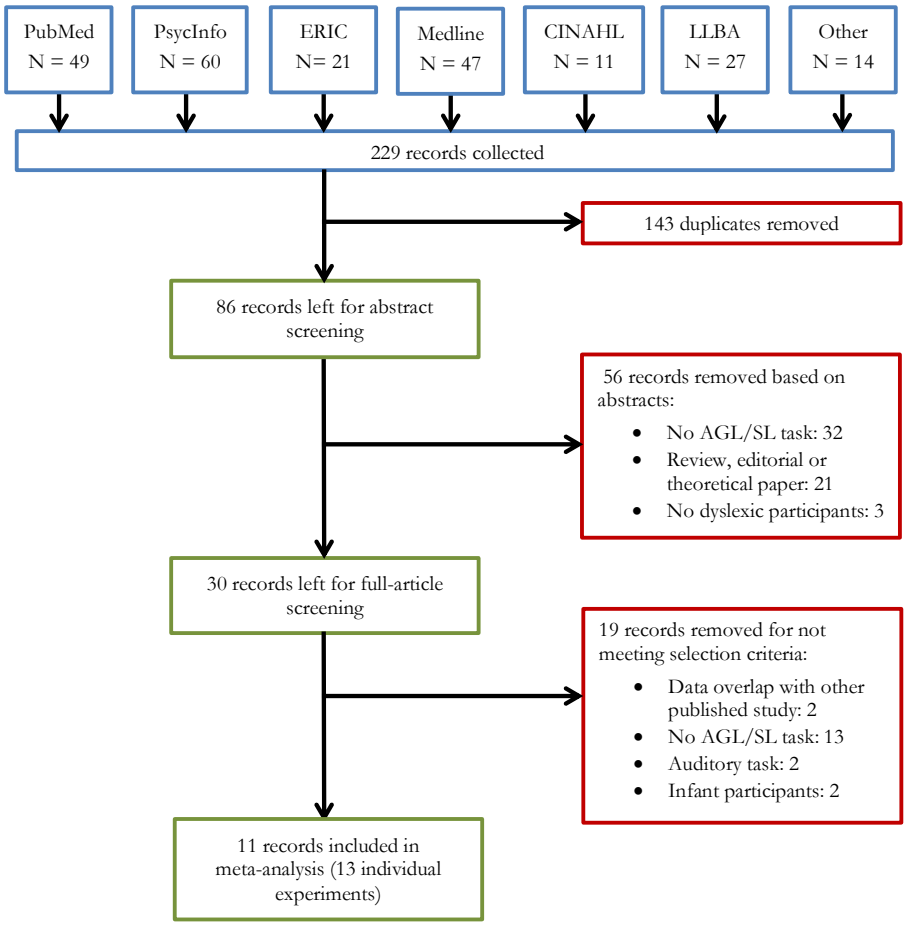


Figure 3.2. PRISMA flow diagram depicting the selection of studies.
Note. The category "other" includes the OATD dissertation database, presentations at conferences, and personal communication. Data overlap: 2 studies that overlapped in data included one bachelor's thesis that contained the same data as a second bachelor's thesis, while the other was Pavlidou's (2010) dissertation of which data was published elsewhere and included in the present meta-analysis.

3.2.3 Data extraction and effect size calculations

The standard measure of learning in an AGL task is the percentage of correct responses (i.e. overall accuracy) during the test phase of the experiment. Therefore, the method for comparing the performance of two groups on an AGL task is to test whether the overall accuracy differs between the study and the control group. In order to calculate a single effect size for each of the included individual studies, the mean, standard deviation (*SD*) and sample size of each of the study groups were extracted from the article. If these data were not available from studies themselves, we asked the authors to supply these. Authors provided these data in three cases (Ise et al., 2012; Laasonen et al., 2014; Samara, 2013), which allowed us to calculate single effect sizes for each individual study.

Calculations were made based on raw data for Rüsseler et al. (2006). For the study by Kahta & Schiff (2016) the mean and 95% confidence interval (CI) had to be gleaned from figures. This was done using DigitizeIt digitizer software (available from http://www.digitizeit.de/). Next, 95% CIs were converted into *SD*s according to Eq. (1), which assumes that authors had computed the CIs with the help of the *t*-distribution. Additionally, the studies by Du (2013), Kahta and Schiff (2016) and Samara (2013) did not report the means and *SD* of the participants' average accuracy scores, but separately the means and *SD*s (or 95% CIs in the case of Kahta & Schiff, 2016) of the participants' percentages of correctly accepted and incorrectly accepted test items (i.e. endorsement rates), which we used to calculate the means and *SD*s of the average accuracy scores. In the absence of data on the correlation (over the participants) between percentage of correctly accepted and correctly rejected test items, and in the absence of good evidence from the literature about what a typical correlation could be, we had to make a conservative estimate of the *SD* of the participants' average percentages. If the correlation is 0, then the variance of the average is smaller than each of the reported variances. If the correlation is 1, then the variance of the average is a weighted average of the two reported variances. The conservative choice for the estimation is therefore to assume that the correlation is 1, so that our estimate of the *SD* of the average of the acceptance and rejection scores is given by (2), where n_1 is the number of correct test items, n_2 is the number of incorrect test items, SD_1 is the observed standard deviation of the correctly accepted percentage, and SD_2 is the observed standard deviation of the correctly rejected percentage.

(1) $SD = \sqrt{n} \times \frac{\text{upper limit} - \text{lower limit}}{\text{t\_crit}\,[n-1]}$[6] (2) $SD_{average} = \sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}}$

For the study by Pothos and Kirk (2004), the *SD*s had to be gleaned from their Figure 4 (p. 71). Additionally, the mean age of participants was not available, but since participants were (under)graduates, most aged between 18 and 30, this study was classified as a study involving adult participants. Appendix F presents an overview of the extracted data that was used for effect size calculation for each included study. Tables 3.1 and 3.2 summarize characteristics of the sample and experimental design of the 13 studies included in the present meta-analysis.

Following data extraction procedures, a single effect size was computed for each individual study, using the *compute.es* package (Del Re, 2014) for *R* software (R Core Team, 2019). In the present meta-analysis, Hedges' *g* effect size[7] and 95% CIs summarize the results from each individual study. Positive Hedges' *g* values indicate that the control group reached higher accuracy levels on the AGL task compared to the group of individuals with dyslexia, whereas negative values indicate the opposite. The 95% CI provides an estimate of the precision of the study's effect size: the larger the CI, the poorer the precision. A combination of the *metafor* (Viechtbauer, 2010) and *meta* packages (Schwarzer, 2012) for *R* software was used to convert the computed individual effect sizes and variances to an average weighted effect size[8] and variance across studies.

For more detail on our meta-analysis and meta-regression techniques, you can access the summarized data, *R* Markdown and HTML files through our project page on the Open Science Framework (OSF): https://osf.io/6qaws/.

---

[6] For large *n*, t_crit [*n*-1] is close to 1.96.

[7] Hedges' *g* is a variation of Cohen's *d* that corrects for biases due to small sample sizes (Hedges, 1981).

[8] A standardized effect size was used as opposed to a raw mean difference score, because the raw mean difference scores and pooled *SD*s of individual studies showed large deviations from the overall raw mean difference score (Bond, Wiitala, & Richard, 2003).

Table 3.1. Overview of study sample characteristics per individual study

Study	Native language	Sample size		Mean age		Matching within studies
		Dyslexia	Control	Dyslexia	Control	
Du (2013)	English	12	12	22.1	22.3	Age, gender, IQ
Ise et al. (2012a)	German	14	17	9.4	9.8	Age, IQ
Ise et al. (2012b)	German	15	15	9.4	9.8	Age, IQ
Kahta & Schiff (2016)	Hebrew	14	15	25.1		MD
Laasonen et al. (2014)	Finnish	36	35	36.1	37.5	Age, gender, IQ, handedness
Nigro et al. (2016a)	Spanish	21	21	8–9	8–9	Age, gender, IQ
Nigro et al. (2016b)	Spanish	21	21	8–9	8–9	Age, gender, IQ
Pavlidou et al. (2009)	English	16	16	10.6	10.6	Age, gender, classroom
Pavlidou et al. (2010)	English	16	16	9.3	9.3	Age, gender, classroom
Pavlidou et al. (2014)	English	16	16	10.3	10.3	Age, gender, classroom
Pothos & Kirk (2004)	English	37	72	MD	MD	MD
Rüsseler et al. (2006)	German	12	12	28.8	32.8	Gender, IQ, handedness
Samara (2013)	English	25	24	21.6	20.5	Age, IQ

Note. Both Ise et al. (2012) and Nigro et al. (2016) reported two separate experiments with distinct participant groups;
MD = missing data.

Table 3.2. Overview of AGL task design per individual study

Study	Stimulus type	Sequence length	Training phase	*N* training sequences	Test phase	*N* test sequences
Child participants						
Ise et al. (2012a)	Letters (CVCV)	5	Active	15	GJ	24
Ise et al. (2012b)	Letters (CCCC)	5	Active	15	GJ	24
Nigro et al. (2016a)	Abstract shapes	4	Passive	36	2-AFC	32
Nigro et al. (2016b)	Letters (CVCV)	4	Passive	36	2-AFC	32
Pavlidou et al. (2009)	Abstract shapes	2–6	Passive	23	GJ	32
Pavlidou et al. (2010)	Abstract shapes	2–6	Active	8	GJ	20
Pavlidou et al. (2014)	Abstract shapes	2–6	Active	8	GJ	20
Adult participants						
Du (2013)	Chinese characters	6–8	Passive	20	GJ	60
Kahta & Schiff (2016)	Letters (CCCC)	5–7	Passive	20	GJ	40
Laasonen et al. (2014)	Abstract shapes	2–6	Passive	23	GJ	32
Pothos & Kirk (2004)	Abstract shapes	2-6	Passive	23	GJ	32
Rüsseler et al. (2006)	Letters (CCCC)	4–7	Active	20	GJ	24
Samara (2013)	Letters (CCCC)	7	Active	12	GJ	48

Note. Stimulus type: several studies used strings that alternated consonants (C) and vowels (V; CVCV), whereas others used consonant strings (CCCC); Training phase: Active = memorization, Passive = exposure; Test phase: GJ = grammaticality judgment, 2-AFC = two-alternative forced-choice

3.3 Results

3.3.1 AGL in dyslexia

Our first goal was to elucidate whether, combining the results from 13 previous studies, individuals with dyslexia perform different from their TD peers on visual AGL tasks. To this end, the effect sizes of all 13 individual studies were combined into a single average weighted effect size using a random-effects model (Hedges & Olkin, 1985). Random-effects models, as opposed to fixed-effect models, allow for variation in true effect sizes between independent studies (Borenstein, Higgins, & Rothstein, 2009). The model was run using the "rma.uni" function in the *metafor* package with the restricted maximum likelihood (REML) method and the adjustment by Knapp and Hartung (2003) for finite numbers of degrees of freedom. Effect sizes for individual studies and the overall average weighted effect size are presented in Figure 3.3. Performance was measured as the overall accuracy score in the test phase of the AGL experiment. Effect sizes ranged from -0.68 to 1.37, with only one effect size in the negative direction (Pothos & Kirk, 2004). All other studies report a lower accuracy level for the group of participants with dyslexia than for the control group. Importantly, as mentioned, some of the individual studies report significant differences, whereas others do not. The meta-analysis reveals that, grouping over 13 studies and despite the negative-estimate study, participants with dyslexia performed significantly worse than control participants (average weighted effect size = 0.46, 95% CI [0.14 … 0.77], p = .008). Looking at studies involving either child or adult participants separately, we find that the average weighted effect size for child studies is significant (N = 7, average weighted effect size = 0.71, 95% CI [0.36 … 1.07], p < .001), whereas it is not for adult studies (N = 6, average weighted effect size = 0.16, 95% CI [-0.36 … +0.69], p = .461). Before investigating whether the observed difference between the adult effect (0.16) and the child effect (0.71) reflects a genuine decreasing difference between dyslexic and non-dyslexic people as a function of age, we inspect the possibility of publication bias.

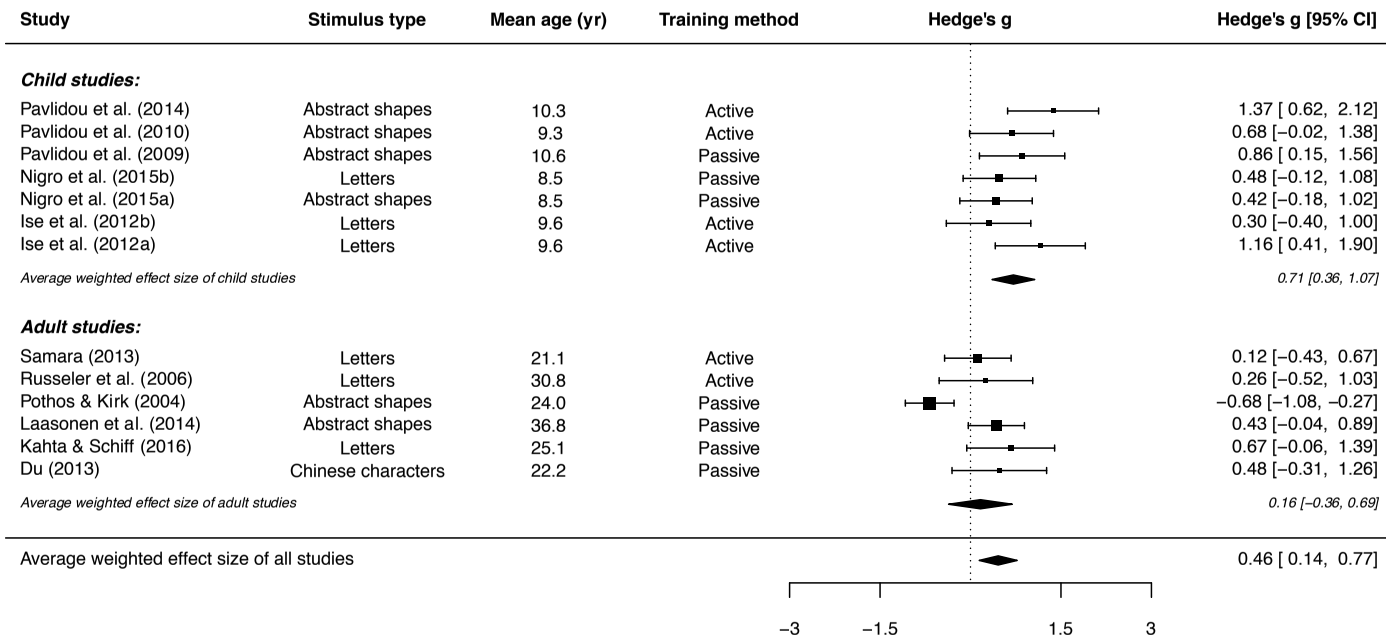


Figure 3.3. Forest plot showing the results of 13 studies investigating AGL performance in individuals with and without dyslexia. The average weighted effect size (and its 95% CI) is indicated at the bottom of the figure.

3.3.2 Publication bias

To verify the interpretability of the abovementioned findings, we examined the possibility of publication bias in our collected sample of studies. This was initially done through examining a standard funnel plot, which plots the standard error (a measure of study precision) against the effect sizes of the individual studies (Figure 3.4, on the left). Generally speaking, in the absence of publication bias, studies should be symmetrically distributed around the average weighted effect size. This distribution takes a funnel shape configuration: studies with high precision are closer to the average weighted effect size, whereas lower precision studies are symmetrically scattered around the average weighted effect size. A linear regression analysis (Egger et al., 1997), using the "metabias" function (Schwarzer, 2012), formally tested the presence of publication bias. It turned out that effect sizes were significantly asymmetrically distributed, skewing to the lower right corner, indicating the presence of a publication bias in our sample (t[11] = 4.014, p = .002).

To evaluate the effect of the publication bias in our sample we approximated what the effect size might be in absence of this bias, using Duval and Tweedie's (2000) trim and fill method ("trimfill" function in the *metaphor* package using the L0 estimator for the number of missing studies). Importantly, the trim and fill method can be used to investigate how sensitive the observed effect is to the presence of potential missing studies, but is not meant as a way to calculate the actual values of missing studies (Duval & Tweedie, 2000; Duval, 2005). By using small studies on the positive side of the funnel plot to impute missing studies on the negative side, the trim and fill method estimated that five studies reporting negative findings are missing in our present sample (Figure 3.4, on the right). When these five imputed missing studies are added to our dataset of 13 studies, the estimated effect size is considerably reduced and is no longer significantly different from zero (average weighted effect size = 0.20, 95% CI [-0.11 … +0.50], p = .205). Note, however, that the trim and fill analysis is known to be a somewhat conservative method for adjusting for publication bias (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Schwarzer, Carpenter, & Rücker, 2010) and the creation of imputed studies can be heavily influenced by a single deviant study, such as the study by Pothos & Kirk (2004) in our sample

(e.g. Borenstein et al., 2009, p. 286).[9] Additionally, this method of adjusting results for publication bias makes the assumption that the asymmetry observed in the funnel plot is caused exclusively by publication bias, while another possible cause for funnel plot asymmetry is heterogeneity between studies (Mavridis & Salanti, 2014). Finally, we cannot be certain that the computed missing studies would indeed have been found in the absence of such a bias (Mavridis & Salanti, 2014). Nonetheless, the results of the present meta-analysis on our selected 13 studies are likely to be overly optimistic in the direction of the existence of the main effect, as the effect can well be nulled by unpublished findings.

3.3.3 Heterogeneity in findings and meta-regression

The second aim of the present study was to explore several factors that may help account for the heterogeneity (between-studies variability) that appears to be present across different studies investigating AGL in participants with dyslexia. Although the main outcome of the present meta-analysis is probably influenced by the observed publication bias, such a bias is less likely to affect meta-regression analyses, which consider secondary effects.

Cochran's Q-test for heterogeneity was significant ($Q[12] = 41.07$, $p < .001$, $I^2 = 71\%$ [0.49 … 0.83]). This result allows us to reject the null hypothesis that all the studies share a common true effect size. As can be seen in Figure 3.3, it appears that some factors may influence the effect size of individual studies. As mentioned, the average weighted effect size for child studies is larger than for adult studies (0.71 for child studies vs. 0.16 for adult studies). Thus, we decided to explore the effect of several potential moderator variables on the effect sizes of individual studies through meta-regression analysis.

---

[9] There exist alternatives to using the L0 method. Using the R0 estimator instead of L0, we find zero missing studies in the present sample, while applying the Copas selection model (Copas 1999; Copas & Shi, 2000) converges to a fully negative CI. Both of these alternative results are due to the presence of the single large study that reports a negative effect (Pothos & Kirk, 2004). The R0 estimate must be incorrect given the significance of the linear regression analysis, and the Copas result must be incorrect because the other 12 included studies show effects in the positive direction.

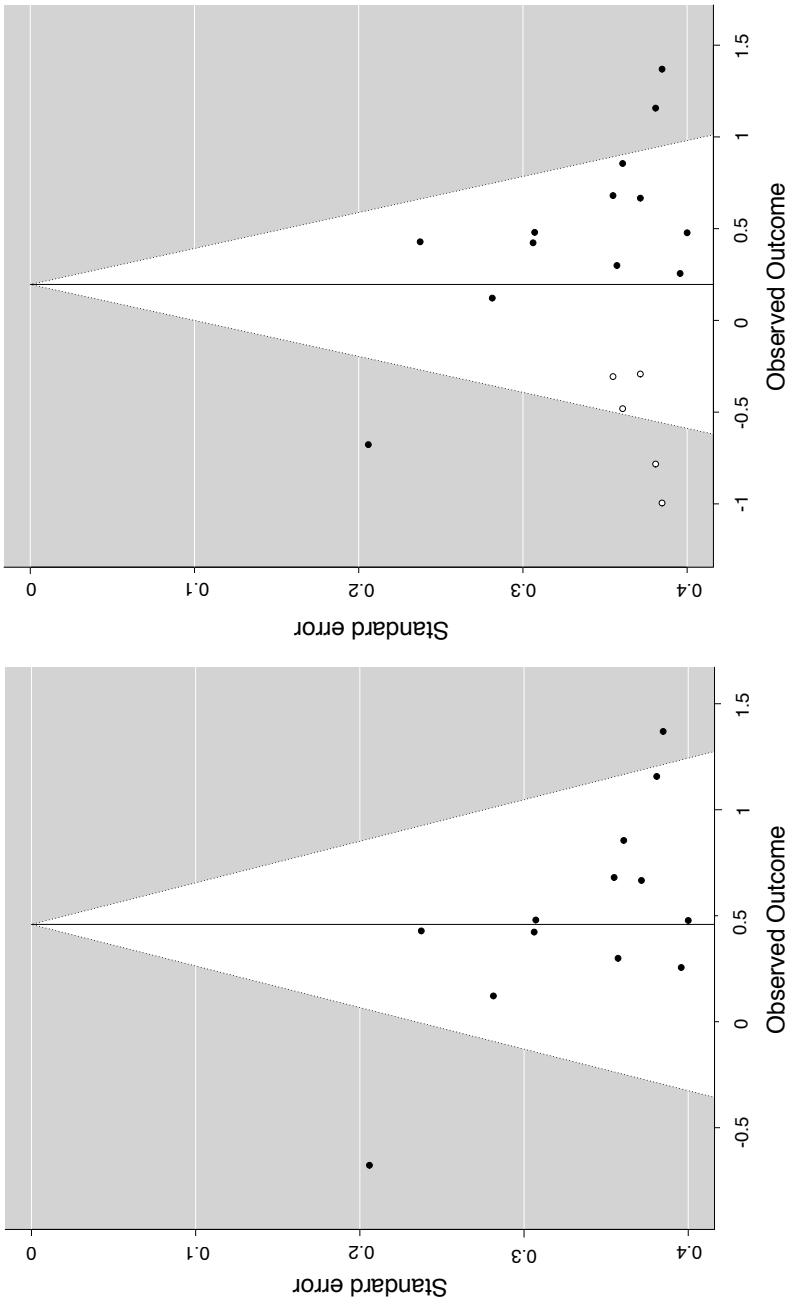


Figure 3.4. Left: Funnel plot showing the relationship between the standard error and the observed outcome (effect size) of individual studies. Right: Funnel plot after controlling for publication bias using the trim and fill method. Solid circles represent included studies, while open circles are imputed missing studies due to publication bias.

Table 3.3. Results from meta-regression analyses exploring the effect of participant and methodological factors on effect sizes of individual studies

Moderator	R^2	Q_{model}	*df*	*p*
1: Age	35%	5.35	1	.041*
2: Training method	1%	1.02	1	.333
3: Stimulus type	0%	0.02	1	.903
4: Age × training method	21%	1.75	3	.659
5: Age × stimulus type	27%	1.94	3	.364
6: Training method × stimulus type	3%	1.04	3	.192

Note. R^2 = the proportion of the total heterogeneity between studies accounted for by the moderator; Q_{model} is the statistic for testing whether the moderator accounts for some of the heterogeneity between studies; *p* is the significance for Q_{model} being greater than *df*.
* $p < .05$

In preparation for the meta-regression analysis, all three binary moderator variables were centered, i.e. coded as: (a) age - $^1/_2$ (child) versus + $^1/_2$ (adult), (b) type of stimulus - $^1/_2$ (abstract shapes) versus + $^1/_2$ (letters), and (c) type of training - $^1/_2$ (passive exposure) versus + $^1/_2$ (active memorization). Random-effects model meta-regression was used to explore the potential value of these factors in explaining variance in effect size between studies. Since the three moderators are correlated, we first tested each of three main effects individually in a separate meta-regression model (Table 3.3). Additionally, we tested each of the three interaction effects individually, in a separate model that included the two relevant main effects (also in Table 3.3). None of the interaction effects turned out to significantly affect the effect sizes of individual studies, so we did not attempt to construct any more complicated models. As shown in Table 3.3, the only model that reaches significance in explaining variance between individual studies is model 1: the main effect of age. This model fits 35% of the heterogeneity (which is greater than 0% with $p = .041$). When studies had adult participants, as opposed to child participants, effect sizes were smaller, reflecting a smaller difference between participants with and without dyslexia. None of the other main or interaction effects were found to significantly fit the heterogeneity between studies. To the extent to which a *p*-value of .041 can be considered statistically significant in this exploration of six possible effects (without correction for multiple testing), we can conclude that the difference between the

observed adult and child effect sizes (0.16 and 0.71) indeed reflects a genuine difference between the two ages in the population.

3.4 Discussion

In the present study, we used meta-analysis and meta-regression to quantitatively review previous research on visual AGL in dyslexia. Our first goal was to elucidate whether the combined findings of thirteen previous studies provide evidence for a difference in visual AGL between individuals with and without dyslexia. The average weighted effect size computed from these individual visual AGL studies, reflecting results from 255 participants with dyslexia and 292 control participants, was found to be moderate and statistically significant. If our 13 selected studies were a sample randomly drawn from an imagined infinite set of possible studies, this finding would indicate that, overall, non-dyslexic people outperform people with dyslexia on visual AGL. Our results would then corroborate the earlier analysis in Schmalz et al. (2017) and strengthen these findings by involving a larger sample of studies (13 instead of 9). Taken together with the meta-analysis of SRT studies by Lum et al. (2013), these results would suggest a general implicit learning deficit in individuals with dyslexia.

Importantly, however, it seems plausible that these results have been influenced by a publication bias in the field of artificial grammar learning in dyslexia (see Schmalz et al., 2017). After conservatively controlling for publication bias, the computed effect size was no longer significant, and the results of the main effect of the present meta-analysis should therefore be regarded as unreliable. Large-scale future studies are needed to confirm the presence of a difference in performance on visual AGL between participants with and without dyslexia.

Extending the previously published meta-analysis by Schmalz et al. (2017) further, the present study aimed to explain the heterogeneity in results of individual studies by investigating the effect of certain methodological variables through a meta-regression analysis. This analysis revealed that the only moderator that (moderately, i.e. without correction for multiple tests) reached significance was the main effect of age: there were smaller differences between dyslexia and control groups for those studies that involved adult participants as

opposed to child participants. This is an indication that the implicit learning deficit might be more pronounced in children with dyslexia than in adults with dyslexia, since similar effects of age have been found in the meta-analysis investigating implicit learning in individuals with dyslexia using the SRT task (Lum, et al., 2013). In line with the interpretation of their results, a possible explanation is that adults make use of compensatory processes (e.g. visual processing, pattern recognition, attentional resources, declarative memory) that enhance performance on visual AGL tasks. Another potential explanation for the age effect lies in the selection of participants. Whereas most studies with adult participants involved university students, child studies selected their participants from a broader population of primary school children. The performance of university students with dyslexia may not be representative of the whole population of adults with dyslexia, as these high-achieving individuals with dyslexia may have more developed compensatory mechanisms. This in turn may result in a smaller difference between the performance of adults with and without dyslexia. We want to note that this effect of age should be interpreted with caution, as it seems to be largely driven by one study that reports better performance in adults with dyslexia than in controls (Pothos & Kirk, 2004, g = -0.68). Thus, future research should examine the possibility of an age effect in visual AGL in dyslexia in further detail by selecting adult participants with dyslexia from all educational levels and comparing them to children on the same visual AGL task.

Although the present meta-analysis suggests that visual artificial grammar learning might be poorer in dyslexia relative to non-dyslexic individuals overall, these results cannot address the issue of causality between implicit statistical learning and literacy skills in this population. Future longitudinal studies are needed to investigate the potential causal link between implicit statistical learning and literacy skills in individuals with and without dyslexia.

Additionally, several factors that could influence the effect sizes of individual studies were not included in the present meta-analysis due to the relatively small number of studies. One such factor is the complexity of the underlying grammar. The level of complexity potentially plays a role in whether participants are able to learn the underlying structure. In fact, a recent meta-analysis of AGL studies with typical populations showed that, indeed, there is a significant correlation between grammar complexity and learners' task performance (Schiff & Katan, 2014). Also related to the difficulty of the task at

hand are factors such as the length of the sequences and the amount of exposure to these sequences. Whereas some studies use a fixed sequence length of 4 (Nigro et al., 2016), 5 (Ise et al., 2012), or 7 (Samara, 2013), other studies use sequences of varying lengths (between 2 and 6 (e.g. Pothos & Kirk, 2004; Pavlidou & Williams, 2014), 4 and 7 (Rüsseler et al., 2006) or 6 and 8 (Du, 2013) individual items). Similarly, whereas some studies include only 69 instances of a grammatical string (e.g. Laasonen et al., 2014; Pavlidou et al., 2009), others include as many as 108 instances (Nigro et al., 2016). Another factor worth investigating is the severity of dyslexia in individual participants, as this may be related to the severity of the deficit in implicit statistical learning. Finally, the modality (visual versus auditory) in which the stimuli are presented may affect the learnability of the grammar for individuals with dyslexia. Future research should investigate the potential effect of the abovementioned factors to gain further understanding of what types of methodological characteristics increase or decrease an AGL task's learnability for individuals with and without dyslexia.

Chapter 4

Statistical learning in dyslexia across three paradigms*

Abstract

Purpose: Statistical learning difficulties have been suggested to contribute to the linguistic and non-linguistic problems observed in children with dyslexia. Indeed, studies have demonstrated that children with dyslexia have problems with statistical learning, but the extent of the problems is unclear. We aimed to examine the performance of children with and without dyslexia across three distinct paradigms using both on- and offline measures, thereby tapping into different aspects of statistical learning.

Methods: 100 children with and without dyslexia (aged 8-11, 50 per group) completed three statistical learning tasks: serial reaction time (SRT), visual statistical learning (VSL), and auditory nonadjacent dependency learning (A-NADL). Learning was measured through online reaction times during exposure in all tasks, and through offline questions in the VSL and A-NADL tasks.

Results and conclusions: We find significant learning effects in all three statistical learning tasks. From this we conclude that, collapsing over groups, children are sensitive to the statistical structures presented in the SRT, VSL and A-NADL tasks. No significant interactions were found between the measures of learning and with group (i.e. dyslexia versus control) in any of the tasks, so we cannot conclude whether or not children with dyslexia perform differently on the statistical learning tasks than their typically developing peers. These results are discussed in light of the proposed statistical learning deficit in dyslexia.

---

* This chapter is a slightly modified version of a published article: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (2019). Statistical learning abilities of children with dyslexia across three experimental paradigms. *PLoS ONE, 14*(8), Article e0220041.

4.1 Introduction

Dyslexia is one of the most common learning disabilities and is characterized by specific difficulties in learning to read and write despite normal intelligence, schooling and socio-economic opportunities and in absence of other impairments (e.g. sensory or neurological impairments; Snowling, 2000). These difficulties in the acquisition of literacy skills are typically associated with problems in related abilities including phonological awareness, lexical retrieval, and verbal short-term memory (e.g. Gathercole, Alloway, Willis, & Adams, 2006; Melby-Lervåg et al., 2012; Ramus et al., 2003). For this reason, the predominant view of dyslexia is that the concomitant reading and writing problems stem from an underlying problem in the processing of phonological information (e.g. Ramus et al., 2003; Snowling, 2001). However, deficits in individuals with dyslexia may include other domains of language (e.g. inflectional morphology and syntax; Rispens & Been, 2007; Waltzman & Cairns, 2000) and non-linguistic cognitive skills such as visual and auditory processing (Stein & Walsh, 1997; Tallal, 2004), attention (Facoetti et al., 2000) and motor functioning (Ramus, 2003; Ramus et al., 2003).

Due to this wide range of observed difficulties, it has been suggested that dyslexia is associated with a domain-general learning deficit rather than a deficit that is specific to the processing of phonological material (e.g. Nicolson & Fawcett, 2007; 2011). This domain-general learning mechanism is often referred to as statistical learning: the ability to extract statistical regularities from sensory input (Frost et al., 2015), which is assumed to be a largely implicit process (e.g. Perruchet & Pacton, 2006). Importantly, statistical learning is put forward as a key ability involved in the acquisition of language and literacy skills as it aids the discovery of the many rules and regularities that are present in spoken and written language (e.g. Arciuli, 2017). In line with this reasoning and the hypothesized statistical learning deficit in dyslexia, evidence shows that statistical learning abilities are related to literacy skills in typical populations. For example, performance on tasks that assess statistical learning abilities has been shown to positively correlate with reading in adults and children (Arciuli & Simpson, 2012) and reading in a second language in adults (Frost et al., 2013). Similarly, children with dyslexia have been shown to perform worse on tasks assessing statistical learning abilities such as the serial reaction time (SRT), auditory statistical

learning (ASL) and artificial grammar learning (AGL) tasks (e.g. SRT: Jiménez-Fernández et al., 2011; Vicari, Marotta, Menghini, Molinari, & Petrosini, 2003; ASL: Gabay et al., 2015; AGL: Pavlidou & Williams, 2014). However, others find no evidence of such an effect (e.g. SRT: Rüsseler et al., 2006; AGL: Rüsseler et al., 2006, Inácio et al., 2018; cued reaction time task: Roodenrys & Dunn, 2008). Literature reviews and meta-analyses have been conducted to investigate the overall group effect in statistical learning studies and have reported significantly poorer performance by individuals with dyslexia as compared to those without dyslexia on both the SRT (Lum et al., 2013) and the AGL overall, although the effect on the AGL may be inflated due to publication bias in the field (Schmalz et al., 2017; van Witteloostuijn et al., 2017, see chapter 3).

The current study aims to investigate to what extent children with dyslexia experience difficulties in the area of statistical learning and to extend recent findings to other statistical learning paradigms. It is important to study children specifically to clarify whether statistical learning principles could potentially be used to improve treatment and clinical outcomes for individuals with dyslexia (see e.g. Plante & Gómez, 2018, on the clinical relevance of statistical learning to children with developmental language disorder [DLD]). Since the hypothesized statistical learning deficit has been claimed to be independent of the domain and modality in which statistical learning is tested, children with dyslexia should experience difficulties across tasks tapping into SL abilities. Therefore, we assess children's statistical learning performance in a range of statistical learning tasks that have previously been shown to be sensitive to learning in (typical) child populations and that span a number of methodological variations of statistical learning tasks (e.g. modality, the type of statistical structure to be learned, online and offline measures): SRT, visual statistical learning (VSL), and auditory nonadjacent dependency learning (A-NADL) tasks. By measuring statistical learning across different experimental paradigms using both online (SRT, VSL, A-NADL) and offline (VSL, A-NADL) measures, and by considering the potential differences in related cognitive abilities including memory and attention, we hope to provide a comprehensive study of statistical learning abilities in children with dyslexia when compared to a control group of age-matched children. Before turning to the methodology of the present study, the following sections present an overview of previous studies investigating statistical learning in dyslexia through the SRT, VSL and A-NADL

paradigms. Subsequently, we discuss several methodological considerations that our design takes into account.

4.1.1 Serial reaction time paradigm

The SRT task measures visuo-motoric sequence learning by exposing participants to a single visual stimulus that repeatedly appears in one of several locations on a computer screen (Nissen & Bullemer, 1987). Without the participants' knowledge, the stimulus follows a predetermined order (i.e. sequence) over three or four locations. During exposure, participants are required to make motor responses that correspond to the locations of the individual stimuli on the screen. As the task unfolds, participants (implicitly) learn the repeated sequence of visual stimuli (locations in array), motor movements, or both, on the basis of the probabilities associated with the sequence. In other words, they learn the probability of the appearance of the stimulus in a given location on the basis of the locations of the previous trials. After participants have been repeatedly exposed to the sequence, they are unknowingly presented with a block of randomly ordered trials. An increase in reaction times (RTs) from predictable (i.e. sequences) to unpredictable (i.e. random) input during exposure is taken as evidence of sensitivity to the sequence presented to them (Nissen & Bullemer, 1987). A range of studies has demonstrated learning in the SRT both in typical adults and in typically developing (TD) children as young as 4 years of age (e.g. Kidd, 2012; Lum, Kidd, Davis, & Conti-Ramsden, 2010).

The SRT task has frequently been used as a measure of statistical learning when investigating group differences between participants with and without dyslexia, both in adult (e.g. Laasonen et al., 2014; Menghini et al., 2010) and child populations (e.g. Deroost et al., 2010; Waber et al., 2003). The difference in sensitivity to SRT structure between participants with and without dyslexia was statistically significant in some studies (e.g. Bussy et al., 2011; He & Tong, 2017, the latter with 40 exposures) but not in others (e.g. He & Tong, 2017; Kelly et al., 2002; Staels & Van den Broek, 2017, the first with 180 exposures). Lum et al. (2013) performed a meta-analysis of 14 such SRT studies involving both adults and children and showed that on average, non-dyslexic people outperform people with dyslexia (weighted average effect size = .449; $p < .001$). To

summarize, the SRT task is known to be sensitive to learning in child populations and has been shown to differentiate between people with and without dyslexia.

4.1.2 Visual statistical learning paradigm

VSL is a paradigm that assesses the capacity for statistical learning by exposing participants to a continuous stream of visual stimuli such as abstract shapes (e.g. Turk-Browne et al., 2005) or cartoonlike figures (e.g. Arciuli & Simpson, 2011; 2012). Unbeknownst to the participants, the stimuli in a VSL task are grouped together in groups of two (i.e. *pairs*) or three (i.e. *triplets*) that always appear together. This task is an adaptation of an auditory task that assesses word segmentation, introduced by Saffran et al. (1996). Thus, in the VSL, the probability of one stimulus following the preceding one differs per trial: while the second (and third) stimulus within a pair (and triplet) is predictable, the first stimulus of the next group is unpredictable. Following repeated exposure to the structured stimuli, a test phase assesses the participants' ability to distinguish previously seen groups of stimuli from groups of stimuli that did not co-occur frequently during exposure. By applying this experimental paradigm, it has been shown that not only adults show sensitivity to this type of statistical structure (Arciuli & Simpson, 2012; Siegelman & Frost, 2015; Turk-Browne et al., 2005), but also school-aged children (Arciuli & Simpson, 2011; 2012; Raviv & Arnon, 2017), as well as infants when tested in a preferential looking time paradigm (e.g. Kirkham et al., 2002). Similar results have been reported for studies involving auditory stimuli including syllables (e.g. Saffran et al., 1996) or non-verbal stimuli such as tones (e.g. Saffran, Johnson, Aslin, & Newport, 1999).

Relevant to the present investigation, only two previous studies have examined the statistical learning abilities of participants with dyslexia using a variant of the VSL task (Sigurdardottir et al., 2017; Singh et al., 2018). In the study by Sigurdardottir et al. (2017), the exposure phase comprised twelve abstract visual shapes that were divided into six pairs of co-occurring stimuli, and participants were subsequently tested in a two-alternative forced-choice (2-AFC) test phase consisting of 72 trials. The results show that adult participants with dyslexia reached lower accuracy levels in the test phase than the control group in the VSL task (68% vs. 78% respectively). The second study investigated the event-related potential (ERP) correlates of statistical learning in children with and

without dyslexia using a visual task (Singh et al., 2018). During the task, children were continuously exposed to series of colored circles and were required to respond to a target color through a button press. Although RT data revealed no difference between children with and without dyslexia (*N* = 8 and 12 respectively), ERP data reveal indications of learning in the control group, but not in participants with dyslexia. Although these studies suggest poorer sensitivity to VSL structures in participants with dyslexia as compared to control participants, no study to date has applied the standard "triplet" paradigm (e.g. Arciuli & Simpson, 2011; 2012; Siegelman & Frost, 2015) to children with dyslexia. Moreover, no data regarding explicit judgments of VSL structure is available on children with dyslexia.

4.1.3 Nonadjacent dependency learning paradigm

Gómez (2002) aimed to test learning of a different type of structure: nonadjacent dependencies in the auditory domain (i.e. A-NADL). In this type of structure, participants learn relationships between nonadjacent elements, ignoring variable intervening elements; for instance, in the string *aXb*, *a* predicts *b* and *X* is a variable intervening element. This experimental design relates to nonadjacent dependencies found in natural language, such as those in inflectional morphology (e.g. *is* eat*ing*, *has* eat*en*, where the auxiliary predicts the inflectional morpheme regardless of the intervening verb; Grama, Kerkhoff, & Wijnen, 2016; Gómez, 2002). Not only adults, but also infants at age 1;6 were sensitive to such nonadjacent dependencies through mere exposure when 24 intervening *X*-elements are used. This is reflected by differences in responses when, after the exposure phase, they are confronted with strings that adhere to the *aXb* grammar as opposed to strings that do not (e.g. *aXc*; Gómez, 2002). However, not much is known about the performance of school-aged children on tasks involving nonadjacent relationships. One previous study has investigated A-NADL in children using the Gómez (2002) design and reports above-chance performance on grammatical items in TD children, suggesting sensitivity to the A-NADL structure (Iao, Ng, Wong, & Lee, 2017).

The same paradigm was used to investigate sensitivity to nonadjacent dependencies in relation to dyslexia. Kerkhoff et al. (2013) tested infants with and without a family risk of dyslexia around the age of 1;6 on a slightly adapted

version of the A-NADL task containing two nonadjacent dependencies of the type *aXb*. In the subsequent test phase that consisted of 8 trials, results revealed a significant interaction between grammaticality and risk group: infants without family risk are sensitive to the A-NADL structure (i.e. they listen longer to ungrammatical than grammatical strings), while infants at risk of dyslexia were less sensitive, if at all. A follow-up study from the same lab examined NADL in the auditory and visual domain in Dutch-speaking adults with and without dyslexia (Kerkhoff, de Bree, & Wijnen, 2017): participants were tested on two versions of the auditory experiment containing either test sentences with familiar *X*-elements or test sentences with novel *X*-elements that aimed to test generalization of the rule. On average, participants were more likely to accept (i.e. endorse) grammatical than ungrammatical sentences in both conditions, reflecting sensitivity to the nonadjacent dependency rule, but no interaction was detected between this measure of learning and group. Similar results are reported for NADL by adults in the visual domain. To summarize, differences in sensitivity to the A-NADL structure were found in infants with and without risk of developing dyslexia, and the results for adults are inconclusive. To our knowledge, no reports of school-aged children with dyslexia on tasks assessing (A-)NADL have been published.

4.1.4 The current study

A number of methodological considerations become apparent from previous literature that are relevant for our investigation of statistical learning in dyslexia. Firstly, and perhaps most importantly, the majority of studies has focused on infant and adult participants. Whereas the SRT and AGL tasks have been used in child populations with and without dyslexia, studies employing alternative paradigms such as the VSL and A-NADL have not been used to investigate statistical learning in school-aged children with dyslexia.

Secondly, although statistical learning is thought to be a domain-general learning mechanism, task parameters and participant characteristics are likely to influence the magnitude of the learning effect found in individual studies (Frost et al., 2015; Siegelman & Frost, 2015). Researchers have previously emphasized the importance of using a range of statistical learning measures within a single sample when investigating the hypothesized statistical learning deficit in children,

as opposed to using only one statistical learning paradigm as is common in most studies (Arciuli & Conway, 2018; West et al., 2017).

Thirdly, VSL and A-NADL tasks have commonly used offline measures to assess learning after exposure. While these measures inform us about the *outcome* of the learning process, they do not inform us about the learning process itself (Kidd et al., 2017; Misyak et al., 2010; Siegelman et al., 2017b; 2018). Recently, two studies have introduced child-friendly VSL (van Witteloostuijn, Lammertink, et al., 2019, see chapter 2) and A-NADL (Lammertink, van Witteloostuijn et al., 2019) tasks that include online measures of learning adapted from previous studies with adult participants (López-Barroso et al., 2016; Siegelman et al., 2018). These online measures reflect participants' sensitivity to statistical regularities during exposure to the stimuli and may provide further insights into the potential differences in performance between children with and without dyslexia when used in addition to the more traditional offline measures.

Finally, studies have shown that performance in statistical learning tasks is affected by cognitive abilities such as attention (e.g. Baker et al., 2004; Toro et al., 2005). Arciuli (2017) has argued that statistical learning is not only related to attention but may also partly rely on (short-term, working and long-term) memory (see also Arciuli & Simpson, 2011; Janacsek & Nemeth, 2015; Lum, Conti-Ramsden, Page, & Ullman, 2012). Important to the present discussion is the fact that individuals with dyslexia have difficulties in the area of attention (e.g. Bosse, Tainturier, & Valdois, 2007; Buchholz & Davies, 2005) and short-term and working memory (e.g. Cowan et al., 2017).

The present study aims to address the abovementioned methodological considerations by assessing the performance of children with and without dyslexia on three different experimental paradigms using a range of online (SRT, VSL and A-NADL) and offline (VSL and A-NADL) measures. In doing so, we want to provide a comprehensive study in which we investigate to what extent children with dyslexia have difficulty in statistical learning. In all analyses, we address two research questions:

1. Do we find evidence of sensitivity to the statistical structure in the SRT, VSL and A-NADL tasks in children overall?
2. Do we find evidence of a difference in performance on the SRT, VSL and A-NADL tasks between children with and without dyslexia?

If children with dyslexia experience general difficulties with statistical learning, we expect to find group differences across the different tasks tapping into statistical learning regardless of the characteristics of the task (e.g. domain, modality or type of structure to be learned). By running subsequent exploratory analyses that control for sustained attention and visual and auditory short-term and working memory, we take into account the possibility that potential group differences in statistical learning are due to underlying differences in these cognitive abilities (i.e. do children with dyslexia experience problems with statistical learning independent of potential difficulties with sustained attention and shot-term and working memory?). Thus, the present study will shed light on the mechanisms underlying the reading problems experienced by individuals with dyslexia: could a domain-general deficit in statistical learning contribute to these problems?

4.2 Materials and methods

4.2.1 Participants

Participants in the present study were tested as part of a larger study that investigates statistical learning and its relationship with language skills in children with dyslexia, children with DLD and TD children. Ten out of 60 participants with a prior formal diagnosis of dyslexia were excluded because they did not meet our pre-determined inclusion criterion of scoring an average of 6 or less (the 10th percentile) on word reading and nonword reading. Similarly, 4 out of 54 children in the TD group were removed for not meeting our inclusion criterion of scoring an average of 8 or more (the 25th percentile). Consequently, the final sample consisted of 50 children with dyslexia (26 girls, 24 boys, age range 8;4 – 11;2, *M* = 9;10) and 50 age-matched TD children (24 girls, 26 boys, age range 8;3 – 11;2, *M* = 9;8). None of the children had diagnoses of (additional) developmental disorders and all children were native speakers of Dutch (at least one parent spoke Dutch at home) and were reported to have IQ levels within the normal range of the general population. Group characteristics, including raw and standardized scores on several background measures, are presented in Table 4.1. It is important to note here that the TD group partly overlaps with studies

investigating statistical learning and its relationship with language in children with DLD (Lammertink et al., 2019a; 2019b; 2020).

Children with dyslexia were recruited through treatment centers in Amsterdam (N = 25) and Amersfoort (N = 10) and through parent support groups on Facebook (N = 11). Four children with dyslexia were tested along with the control group in four schools across the province of Noord-Holland in the Netherlands. The ethical committee of the Faculty of Humanities of the University of Amsterdam approved the protocol for the present study in 2016. All parents and/or legal guardians of participants were informed about the project through a newsletter. Compliant with the regulations of the ethical committee, informed consent was obtained from the parents and/or legal guardians of children with dyslexia prior to testing (active consent). For the control group, schools and teachers consented to participation, and parents and/or legal guardians could retract permission of including their child up to 8 days following testing (passive consent).

To compare the group of participants with dyslexia with their TD peers on the range of included background measures, we fitted linear models on the raw data using the "lm" function for *R* software (R Core Team, 2019). No significant differences were found between the chronological ages of the groups (t = 0.839, p = .40), the groups' socio-economic status (SES; t = 0.173, p = .86) or non-verbal reasoning (t = -0.041, p = .97). SES scores were obtained from the *Netherlands Institute for Social Research* (NISR) on the basis of children's home or school postal codes depending on the testing location. These SES scores were calculated by the NISR in 2017 and indicate the social status of a given neighborhood in comparison to other neighborhoods in the Netherlands (open source data that can be accessed through the NISR website). Non-verbal reasoning was assessed through *Raven's Standard Progressive Matrices* (Raven & Raven, 2003). We also measured children's reading of single Dutch words (*Een Minuut Test*; Brus & Voeten, 1972) and pseudo-words (*Klepel*; van den Bos, Spelberg, Scheepsma, & de Vries, 1994), their spelling (*Schoolvaardigheidstoets Spelling*; Braams & de Vos, 2015) and their rapid automatized naming (RAN) of pictures and letters (*Continu Benoemen en Woorden Lezen*; van den Bos & Lutje Spelberg, 2007). In line with expectations, analyses show that children with dyslexia performed significantly more poorly than the TD children on all measures assessing literacy skills (reading words: t = -13.83, p = $9 \cdot 10^{-25}$, reading

pseudo-words: t = -16.75, p = $1.7 \cdot 10^{-30}$, spelling: t = -11.42, p = $9.4 \cdot 10^{-20}$, RAN pictures and letters: t = -4.985, p = $2.7 \cdot 10^{-6}$ and t = -5.421, p = $4.3 \cdot 10^{-7}$ respectively).

We assessed cognitive abilities that are often associated with statistical learning and that may differ between our groups of participants with and without dyslexia: short-term and working memory and attention (see Table 4.1). Short-term and working memory were tested in the auditory domain with the forward and backward digit span tasks from the Dutch version of the *Clinical Evaluation of Language Fundamentals* (CELF; Kort, Schittekatte, & Compaan, 2008) and using forward and backward versions of the dot matrix task in the visuospatial domain (Alloway, 2012). Note that, for the dot matrix task, standardized scores are unavailable and data is based on 49 children with dyslexia, due to missing data for one participant as a result of equipment failure. Sustained attention was measured through the *Score!* subtest of the Dutch *Test of Everyday Attention for Children* (TEA-Ch; Schittekatte, Groenvynck, Fontaine, & Dekker, 2007). In this task, children perform 10 items that contain between 9 and 15 target sounds that are presented at varied intervals. Their task is to silently count the target sounds, reflecting the child's ability to maintain attention over time. The digit span backward and dot matrix forward and backward did not reveal significant differences between participants with and without dyslexia (digit span backward: t = -1.257, p = .21, dot matrix forward: t = -0.667, p = .51, dot matrix backward: t = -1.248, p = .22). Digit span forward performance (i.e. verbal short-term memory) was significantly poorer in participants with dyslexia as compared to their TD peers (t = -5.36, p = $5.5 \cdot 10^{-7}$). The groups differed marginally significantly in sustained attention (t = -1.939, p = .055). Given these findings, we explore whether adding the digit span forward and sustained attention scores to our models influences our findings regarding statistical learning performance (see §4.2.6 on scoring and analysis).

Table 4.1. Minimum, maximum and mean (*SD*) raw and standardized scores on background, memory and sustained attention measures.

	Dyslexia (*N* = 50)		**Control (*N* = 50)**	
	Raw	Standardized	Raw	Standardized
Age	8;4 – 11;2	N/A	8;3 – 11;2	N/A
	9;10 (0;9)		9;8 (0;10)	
SES	-3.31 – 2.09	N/A	-1.28 – 1.41	N/A
	0.2 (1.2)		0.2 (1.1)	
Nonverbal reasoning[a]	23 – 49	7 – 95	16 – 55	6 – 98
	37.2 (6.6)	55.7 (25.0)	37.3 (8.1)	60.1 (28.1)
Reading words[b]	8 – 59	1 – 7	44 – 92	7 – 15
	34.1 (11.7)	3.3 (2.1)	66.3 (11.6)	10.5 (2.2)
Reading pseudo-words[b]	8 – 39	1 – 7	33 – 89	7 – 15
	22.0 (8.0)	4.4 (1.6)	61.0 (14.4)	11.1 (2.2)
Spelling[a]	0 – 17	0 – 71	9 – 27	6 – 95
	8.4 (4.6)	11.8 (13.7)	18.6 (4.7)	49.9 (24.7)
RAN pictures[b]	35 – 80	2 – 14	30 – 63	5 – 16
	53.2 (10.2)	7.7 (2.7)	44.1 (7.3)	10.7 (2.8)
RAN letters[b]	23 – 79	1 – 12	18 – 46	3 – 16
	36.1 (10.4)	5.4 (2.7)	27.2 (5.5)	9.6 (3.1)
Sustained attention[b]	1 – 10	1 – 13	3 – 10	3 – 14
	7.0 (2.5)	7.4 (3.3)	7.8 (1.8)	9.1 (3.0)
Digit span forward[b]	4 – 11	1 – 13	6 – 12	5 – 15
	7.3 (1.5)	7.7 (2.6)	8.9 (1.5)	10.7 (2.9)
Digit span backward[b]	2 – 7	1 – 14	2 – 8	4 – 16
	4.2 (1.1)	9.0 (2.5)	4.5 (1.5)	10.0 (3.2)
Dot matrix forward	15 – 35	N/A	13 – 34	N/A
	25.1 (4.7)		25.7 (5.1)	
Dot matrix backward	8 – 35	N/A	15 – 34	N/A
	22.9 (5.0)		24.1 (4.9)	

Note. Raw scores: number correct on the Raven (max = 60), number words and pseudo-words read correctly within 1 minute and 2 minutes respectively, number of words spelled correctly (max = 30), number of seconds spent on the task in case of the RAN (i.e. higher score = lower performance), number correct on sustained attention (max = 10), digit span (max = 16), and dot matrix (max = 36). Standardized scores: [a] percentile scores (norm = 50) or [b] norm scores (norm = 10).

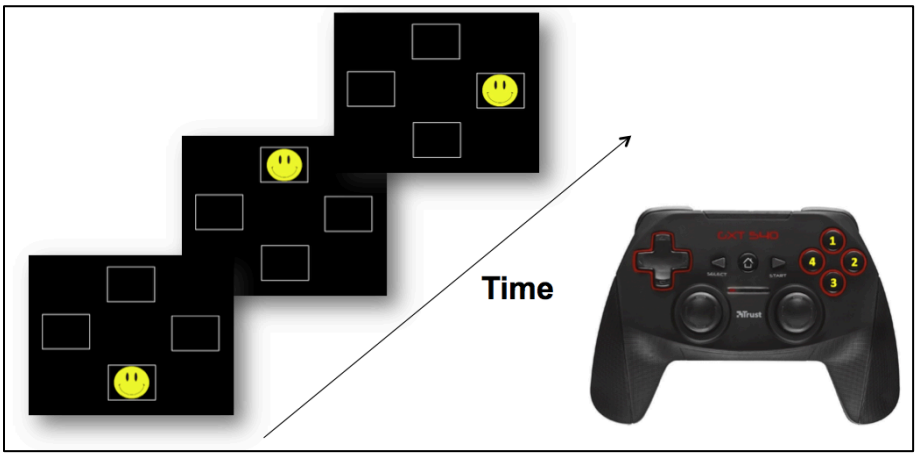


Figure 4.1. SRT task, set up of the experiment. Left: a yellow smiley appeared in one of four marked locations on a tablet screen. Right: participants were required to press the corresponding buttons on a gamepad controller.

4.2.2 SRT task

A visual stimulus (yellow smiley face) repeatedly appeared in one out of four marked locations on a black background presented on a tablet screen. Participants were instructed to press corresponding buttons on a gamepad as quickly and accurately as possible and practiced the task in 28 trials (see Figure 4.1). Each instance of the visual stimulus was visible until a response was given, with a 250 milliseconds interval before the next instance of the stimulus appeared. Participants had a maximum of 3 seconds to respond before the task would move on to the next instantiation of the stimulus automatically.

Unbeknownst to the participant, the stream of stimuli was divided into seven underlying blocks. The first block contained 20 random trials. Blocks 2 through 5 and block 7 contained structured input that consisted of six repetitions of a 10-item sequence (i.e. sequence blocks, 60 trials each). The sequence consisted of a constant order of locations (quadrants) in which the visual stimulus appeared (quadrants 4, 2, 3, 1, 2, 4, 3, 1, 4, 3). In disruption block 6, the appearances of the stimulus no longer followed the sequence, but was presented in random order (i.e. 60 random trials). Both accuracy and RT to each stimulus presentation were recorded. If learning takes place in the SRT task, RTs to predictable input averaged over sequence blocks 5 and 7 are expected to be

shorter than RTs to unpredictable input in the intervening disruption block (Nissen & Bullemer, 1987). The SRT task in the present study did not include an explicit offline test phase.

4.2.3 VSL task

4.2.3.1 VSL task: online exposure phase

In the VSL task, visual stimuli were presented one at a time in the middle of a tablet screen. Without the participants' knowledge, stimuli appeared in the same four groups of three (i.e. triplets; *ABC*, *DEF*, *GHI*, and *JKL*). The exposure phase of the VSL task is divided into four blocks containing six repetitions per triplet, resulting in 24 repetitions of each triplet. Following previous studies adopting a similar structure (Arciuli & Simpson, 2011; 2012; Turk-Browne et al., 2005), triplets could not appear twice in a row and pairs of triplets could not be repeated (i.e. sequences such as *ABC*, *ABC* or *ABC*, *JKL*, *ABC*, *JKL* could not occur). The VSL structure can be expressed in terms of predictability through the TPs (the probability of event i+1 given event i): given the occurrence of element *A*, the TP to element *B* is 1 and the same holds for element *C* given element *B*. The TP when crossing a triplet boundary is low. Thus, whereas elements 2 and 3 within triplets (e.g. stimuli *B* and *C* in the triplet *ABC*) are completely predictable, the first element of the following triplet (e.g. stimulus *D* of the triplet *DEF*) is less predictable.

The self-paced nature of the task entails that participants responded to each individual stimulus by pressing the space bar, upon which the next stimulus appeared after 200 milliseconds (Siegelman et al., 2018; van Witteloostuijn, Lammertink et al., 2019, see chapter 2). We recorded RTs to individual stimuli, which were used as an online measure of learning. If learning takes place, RTs to predictable stimuli (i.e. elements 2 and 3 within triplets) are expected to be shorter than RTs to unpredictable stimuli (i.e. element 1 within triplets). Thus, learning in the online phase of the VSL is reflected by a difference in RTs to predictable as compared to unpredictable stimuli, since sensitivity to the statistical structure is hypothesized to result in faster processing of predictable stimuli (as in the SRT task).

As part of the cover task, three stimuli per block were presented twice in succession (i.e. 12 repetitions in total; Arciuli & Simpson, 2011; 2012). In the event of a repeated stimulus, participants were required to respond by tapping the alien on the touch screen. Each triplet contained a double stimulus three times throughout the exposure phase, all three elements within the triplet once (e.g. the triplet *ABC* occurs once as *AABC*, *ABBC*, and *ABCC*). In each block, three distinct triplets contained a double stimulus in random positions of the stream of stimuli, again all three element positions within triplets once (e.g. *AABC*, *DEEF*, *GHII*).

4.2.3.2 VSL task: offline test phase

To test participants on their acquired knowledge of the triplet structure, they were tested in an offline test phase subsequent to exposure that consisted of 40 multiple-choice questions. Using the same set of 12 stimuli as used in the familiarization phase, four foil triplets were created (*AEI*, *DHL*, *GKC*, and *JBF*, all with TPs of 0 within triplets). Participants first received three-alternative forced choice (3-AFC) questions in which they were asked to complete a missing shape (N = 16, chance level = $^1/_3$) and subsequently questions in which they were required to pick the more familiar pattern out of two options (2-AFC; N = 24, chance level = $^1/_2$). Test items either tested complete triplets (3-AFC: N = 8, 2-AFC: N = 8) or pairs within triplets (3-AFC: N = 8, 2-AFC: N = 16). Learning in the VSL test phase is evidenced by above-chance performance, since above-chance performance reflects participants' ability to explicitly judge which patterns belong to the statistical structure in the VSL task.

4.2.3.3 VSL task: procedures

Importantly, the effect of single stimuli or triplets and the effect of the order of appearance during familiarization and testing were counter-balanced: two sets of triplets (and foil triplets) were created using the same set of 12 stimuli, and two random orders of the presentations of triplets during exposure and testing were created. This resulted in four versions of the experiment, to which participants were randomly assigned.

Before the exposure phase, participants performed two practice phases consisting of an alternative set of stimuli. In the first practice phase, participants practiced sending home the aliens by pressing the space bar (N = 16). In the second part, children were instructed to pay attention to double stimuli and instructed to tap the touch screen in these cases (N = 18, 3 double stimuli). Importantly, children were instructed to pay attention to the aliens and were informed that some of the aliens liked each other and stood in line together. In between the four blocks of the experiment, participants received stickers for a diploma and were stimulated to pay attention to the aliens. Prior to each of the two parts of the offline test phase, participants received instructions and a practice trial during which they were encouraged to make a guess in case they were unsure of the correct response.

4.2.4 A-NADL task

4.2.4.1 A-NADL task: online exposure phase

Children were exposed to an artificial language that, unbeknownst to them, contained two nonadjacent dependencies in 80% of the trials: *tep X lut* and *sot X mip*, where *tep* predicted *lut* and *sot* predicted *mip* and the variable intervening *X*-element always consisted of two syllables (e.g. *wadim*, N = 24; e.g. Gómez, 2002). The remaining 20% of the trials were filler trials that deviated from the two nonadjacent dependencies. These trials can be described as *fXf* trials and resembled the *aXb* nonadjacent dependency structure: the elements in the *f* positions consisted of one-syllable nonwords (N = 24) and were separated by the same *X*-elements used in the nonadjacent dependencies. Appendix G presents an overview of all X- and *f*-elements used in the present experiment (adapted from Lammertink et al., 2019a, Table 2, p. 12). In filler trials, however, the first *f*-element did not predict the second *f*-element. Each trial in the A-NADL thus consisted of three elements and was between 2067 and 2908 milliseconds long (M = 2415 milliseconds) with an interval of 250 milliseconds between elements. All stimuli used in the A-NADL were created in accordance with Dutch phonotactic constraints, followed a natural Dutch sentence prosody (e.g. *het meisje loopt*, the girl walks), and were recorded by a female native speaker of Dutch.

The task consisted of a total of 270 trials divided into five blocks: four blocks containing the nonadjacent dependency rules and fillers (rule blocks 1 – 3 and 5) and one intervening block in which the strings did not contain the nonadjacent dependency rules (disruption block 4). Forty-eight trials in each of the rule blocks contained the two nonadjacent dependencies (24 times *tep X lut* and 24 times *sot X mip*) in addition to 12 fillers, resulting in a total of 60 trials per rule block. Both nonadjacent dependencies were presented in combination with each *X*-element once in each block and thus repeated four times during the exposure phase (i.e. 96 exposures to each nonadjacent dependency). The disruption block contained 30 trials in which the rule structure was disrupted: trials were of the structure *f X lut* and *f X mip*, so that the occurrences of *lut* and *mip* were no longer predictable (*N* = 12 each). The remaining six trials were filler items. Combinations of filler elements (*f*) and *X*-elements were unique and only appeared once across the duration of the exposure phase (*N* = 54).

Importantly, the online measure of learning was a word-monitoring task that required participants to attend to the speech stream and track the occurrence of a "target" (i.e. a specific nonword) and respond as quickly as possible by pressing a button on an external button box (Lammertink, van Witteloostuijn et al., 2019; López-Barroso et al, 2016). The target was always the predictable *b* element of one of the two nonadjacent dependencies (i.e. *lut* or *mip*) and participants were randomly assigned to one of two experiment versions (version 1: target = *lut*, version 2: target = *mip*). Predictable element *b* of the unattended nonadjacent dependency will henceforth be referred to as the "nontarget" (version 1: nontarget = *mip*, version 2: nontarget = *lut*). When the trial contained the target, participants were required to press the green button, while they were required to press the red button when the trial did not contain the target. For example, in version 1 of the experiment, where *lut* was the target, participants had to press the green button when trials contained the target word *lut* (rule blocks: *tep X lut*, disruption block: *f X lut*) and press the red button when trials contained the nontarget word *mip* (rule blocks: *sot X mip*, disruption block: *f X mip*) or contained neither *lut* or *mip* as was the case in filler items. Children had 1500 milliseconds to respond to each trial before the experiment moved on to the next trial automatically. Accuracy and RT were recorded for each individual trial.

As in the SRT and VSL tasks, learning in the online measure of the A-NADL is defined as the difference in RTs between predictable and unpredictable

input. The target and non-target words were predictable during rule blocks (they were always preceded by the corresponding *a* element as in *tep X lut* and *sot X mip*) but were no longer predictable in the disruption block (they were no longer preceded by the *a* element but by a variable *f* element as in *f X lut* and *f X mip*). Thus, mimicking the structure of the SRT task, learning in the A-NADL task is evidenced by shorter RTs to both target and nontarget trials in rule blocks 3 and 5 as opposed to the intervening disruption block.

4.2.4.2 A-NADL task: offline test phase

Participants were tested on their acquired knowledge of the nonadjacent structure through an offline grammaticality judgment task (GJT; $N = 16$). They were required to indicate whether they had previously heard each string by saying either "yes" (endorsement) or "no" (rejection). Eight items were grammatical strings (e.g. *sot densim mip*) and eight were ungrammatical strings where the nonadjacent dependency structure was disrupted (e.g. *sot filka lut*). Similarly, eight strings contained familiar *X*-elements used during exposure and eight strings contained novel *X*-elements that were only used during the test phase. Two additional items that contained three *X*-elements (i.e. *XXX*) functioned as filler items and were not included in the analyses. If learning in the test phase of the A-NADL is successful, we expect to find a higher proportion of endorsements as opposed to rejections to grammatical strings than to ungrammatical strings. A visual representation of the on- and offline phases of the A-NADL used in the present study is provided in Appendix H (adapted from Lammertink et al., 2019a, Figure 1, p. 12).

4.2.4.3 A-NADL task: procedure

There were two counterbalancing variables in the A-NADL: children either received a version of the task where the target was *lut* or the target was *mip* and the location of the green and red buttons on the external button box were counter-balanced. Participants were randomly assigned to the four versions of the experiment.

Participants were seated behind a tablet and held the button box in their hands, using both thumbs to press the buttons. The auditory stimuli were played

through headphones. The word-monitoring task was framed as a game in which the participant helped a monkey to pick bananas. Children were told that they would hear three-word sentences and had to press the green button when they heard the target and the red button when they did not. In accurate trials, the monkey was rewarded with a banana. Children were instructed to pay attention to all three words in the sentences, since they would receive questions at the end. A practice phase containing six items preceded the start of the experiment, which was repeated until they reached a score of 4 out of 6 correct (for which they had to master the motorics and press in time). During the exposure phase, the experiment was broken up into short blocks containing 30 trials each. Following these blocks, children received feedback on the number of bananas they picked and received a sticker for their diploma. Subsequent to exposure, the experimenter instructed the participant that they would hear sentences one at a time and to indicate whether they had heard the sentence before or not. Two practice items preceded the GJT and children were encouraged to guess if they were uncertain of the answer.

4.2.5 General procedure

All statistical learning tasks were programmed and ran using E-prime 2.0 software (Psychology Software Tools, 2012; Schneider et al., 2012) on a Windows Surface 3 tablet with touchscreen and keyboard. Auditory instructions (and stimuli in the case of the A-NADL) were played over Sennheiser HD 201 headphones. Additional materials included the gamepad used in the SRT task (Trust wired gamepad GXT540) and the external button box used in the A-NADL task.

As mentioned previously, participants in the present study were tested as part of a larger study. Children were tested individually by an experimenter in a quiet room either at home or at school. Testing lasted approximately three hours, divided over three testing sessions that lasted around an hour. In each of these sessions, one of the statistical learning tasks was administered along with three or four of our linguistic or cognitive measures (each of these was measured only once). The order of the sessions (and the order of tasks within sessions) was counter-balanced: six testing orders were created to which participants were assigned randomly. Thus, the order of the statistical learning tasks (order 1: A-

NADL, SRT, VSL; order 2: SRT, VSL, A-NADL; order 3: VSL, A-NADL, SRT), as well as the linguistic and cognitive measures, was semi-randomized. Each child was rewarded for their participation with stickers on a diploma and a small present after completing the three sessions.

4.2.6 Scoring and analyses

Online RT data of the SRT, VSL and A-NADL tasks was analyzed with linear mixed effects models that were built using the *lme4* package (version 1.1-13; Bates et al., 2014) for *R* software. Similarly, the *lme4* package for *R* was used to build generalized linear mixed effects models for the offline accuracy data in the VSL and A-NADL tasks. Wherever possible, a confidence interval (CI) was computed by the profile method (*stats* package version 3.5.2 for *R* software; R Development Core Team, 2008), and a corresponding *p*-value was obtained by interpolation among the profiles for different CI criteria (e.g. a *p*-value of .03 was concluded if one of the edges of the 97 percent CI was zero; see "get.p.value" function on the open science framework [OSF]; link provided below). In the A-NADL offline measure, some CIs were computed using Wald's approximation for CI's and *p*-values are obtained from the model output. This was only done when (1) the profile method failed to provide CIs, *and* (2) we did not want to further decrease the random effects structure, *and* (3) the result was non-significant.

For all analyses, continuous predictors were centered and scaled, while categorical predictors were coded into orthogonal contrasts. Group is always orthogonally coded such that the control group is marked as $-1/2$ and the dyslexia group is marked as $+1/2$. Therefore, the effect of group is always interpreted as the change in effect when moving from the control group to the group of participants with dyslexia (see Table 4.2 for an overview of all orthogonally coded categorical predictors per statistical learning task and following sections for further explanation). In line with Barr et al. (2013), models contained the maximal random effect structure, unless this resulted in a failure to fit the model or in (near-)perfect correlations between the random effects in which case reductions were performed that are explicitly justified in the text. Raw data and *R* Markdown and html files detailing all analyses of the SRT, VSL and NADL tasks can be accessed through the following link to our OSF project page: https://osf.io/t8scv/.

For online RT measures, analyses were first run on the raw data. However, in all three tasks, this resulted in non-normally distributed residuals of the linear mixed effects models as evidenced by their Quantile–Quantile ("QQ") plots. Therefore, we decided to use a rank order transformation, which is a principled non-arbitrary way to reduce the effect of outliers and to reduce skewness in the distribution of the residuals (see Baguley, 2012, p. 354–358). The commonly used log-transformation was not appropriate due to the presence of negative RTs. The rank-order transformation was done by ranking the *N* pieces of pooled data from 1 to *N*, then computing the inverse cumulative Gaussian distribution (with the following formula in *R*: qnorm[(ranking-0.5)/*N*]); the model estimates hereby come to represent differences in z-values (Δz; e.g. the main effect of a binary predictor is given by the change in z-value from one level to the other).

As part of our exploratory analyses, we compute additional models for each of our statistical learning measures to investigate the effect of adding sustained attention and verbal short-term memory as continuous predictors; the fit of each of these models is compared statistically to the model without these two measures. At the request of reviewers, children's chronological age was added as an exploratory (continuous) predictor in all models. This was done to reduce variance and to examine whether age interacts with the measures of learning in the statistical learning measures (relating to research question one) and group (relating to research question two). Only significant findings regarding the exploratory effect of age are included in the results section. The subsequent sections provide further details regarding the pre-processing of the data and the analyses of the three statistical learning tasks.

Table 4.2. Orthogonal contrast coding of categorical predictors in the SRT, VSL and A-NADL tasks.

	Predictor	Contrast coding	Purpose
SRT	Block (Bl)	1: Bl 6 = - $^2/_3$, Bl 5 and 7 = + $^1/_3$	RQ 1
		2: Bl 5 = - $^1/_2$, Bl 7 = + $^1/_2$	Exploratory
	Group	TD = - $^1/_2$, DD = + $^1/_2$	RQ 2
VSL	Element (El)	1: El 1 = - $^2/_3$, El 2 and 3 = + $^1/_3$	RQ 1
		2: El 2 = - $^1/_2$, El 3 = + $^1/_2$	Exploratory
	Group	TD = - $^1/_2$, DD = + $^1/_2$	RQ 2
	Triplet Set (TS)	TS A = - $^1/_2$, TS B = + $^1/_2$	Exploratory
	Random Order (RO)	RO 1 = - $^1/_2$, RO 2 = + $^1/_2$	Exploratory
A-NADL	Block (Bl)	1: Bl 4 = - $^2/_3$, Bl 3 and 5 = + $^1/_3$	RQ 1
		2: Bl 3 = - $^1/_2$, Bl 5 = + $^1/_2$	Exploratory
	Grammatical	No = - $^1/_2$, Yes = + $^1/_2$	RQ 1
	Group	TD = - $^1/_2$, DD = + $^1/_2$	RQ 2
	Generalization	No = - $^1/_2$, Yes = + $^1/_2$	Exploratory
	Target Type	Target = - $^1/_2$, Non-target = + $^1/_2$	Exploratory
	Experiment Version	Lut = - $^1/_2$, Mip = + $^1/_2$	Exploratory

Note. RQ = research question: RQ 1 pertains to the overall learning effect, while RQ 2 regards the effect of group (dyslexia versus control) when looked at in interaction with the effect of learning overall. Exploratory predictors and contrasts are included either because predictors are counter-balancing factors or because predictors need to be orthogonally coded (i.e. in the case of predictors with two contrast codings).

4.2.6.1 SRT task

The first block of the SRT, containing 20 random presentations of stimuli, was removed from analysis. Furthermore, incorrect responses and trials in which no response was given were removed from the data file (5.9% data loss).

The linear mixed effects model was run using normalized RTs as the dependent variable. Since online sensitivity to the sequence in the SRT task is measured as the difference in RTs to predictable versus unpredictable input, our analysis contrasted RTs in sequence blocks with RTs in the intervening disruption block that contained random input in order to answer research question one (i.e. within-participant predictor Block: block 5 and 7 vs. block 6).

The categorical predictor Block was orthogonally coded into two contrasts: the effect of learning (i.e. random block 6 coded as - $^2/_3$ vs. sequence blocks 5 and 7 coded as + $^1/_3$ each, thereby comparing random block 6 to the average of sequence blocks 5 and 7) and the contrast between the two sequence blocks (block 5 vs. block 7 coded as - $^1/_2$ and + $^1/_2$ respectively). Further predictors in the model included the between-participants predictors Group (control versus dyslexia) and Age. To answer our second research question, we looked at the interaction between the first level of Block and Group. The model included by-subject random intercepts and by-subject random slopes for Block.

4.2.6.2 VSL task

Our scoring and analysis procedures of the online RT measure followed those by van Witteloostuijn, Lammertink et al. (2019, see chapter 2). The RTs to the first triplet in each block of the experiment were removed (4.2% data loss). This was done because these responses are likely to deviate from participants' normal patterns. Additionally, RTs shorter than 50 milliseconds were removed from the dataset, as these were assumed to reflect cases in which the participant did not process the stimulus (0.2% data loss).

Sensitivity to the structure is measured as the difference in RT to unpredictable versus predictable elements within triplets; this sensitivity may depend on time (research question one). Thus, the model fitted normalized RTs as a function of the within-participant predictors Element (element 1, 2 and 3 within triplets) and Time (repetitions 1-24 of triplets). The categorical predictor Element was orthogonally coded into two contrasts: the effect of learning (i.e. element 1 coded as - $^2/_3$ vs. element 2 and 3 coded as + $^1/_3$ each), and the contrast between the two predictable elements (element 2 coded as - $^1/_2$ and element 3 coded as + $^1/_2$). The interaction between the effect of learning (i.e. the first level of Element) and the between-participant predictor Group (control versus dyslexia), and its three-way interaction with Group and Time were of interest to our second research question. Two counter-balancing factors were included in the model as within-participant predictors (Triplet Set A and B coded as - $^1/_2$ and + $^1/_2$ respectively and Random Order version 1 and 2 also coded as - $^1/_2$ and + $^1/_2$ respectively). Finally, the model contained the exploratory between-participants predictor Age. The random effect structure included by-subject and

by-item intercepts, as well as by-subject random slopes for Element and Time and the interaction between the two. The individual aliens used in the experiment (N = 12) were used for the random intercepts for item. By-item random slopes for group were removed, since these resulted in a perfect correlation between the random intercept for item and the by-item random slopes for group (i.e. the model was overparameterized). This removal did not result in a decrease in the fit of the model (χ^2 = 0.0655, df = 2, p = .968). In order to compute the CIs and p-values of the final model, the interaction between Element and Time, which was non-significant, was removed from the random effects structure.

In the offline test phase, responses were coded as either correct or incorrect (i.e. 1 or 0). Accuracy is expressed as the proportion of correct responses, such that chance levels are $^1/_3$ and $^1/_2$ for the 3-AFC and 2-AFC questions respectively. No accuracy data was removed prior to running the generalized linear mixed effects models.

Two models were constructed to analyze the 3-AFC and 2-AFC accuracy data separately. To answer our first research question as to whether learning took place, we examined whether the proportion of accurate responses exceeded chance level, which is reflected in the intercept of the generalized linear mixed effects models (if performance is significantly above chance level, the CI does not contain the chance level probability associated with that task). As for the second research question, the model contained the between-participants predictor Group (control versus dyslexia). Following the structure of the online VSL model, the offline models further contained the orthogonally coded Triplet Set and Random Order as within-participant predictors, Age as an exploratory between-participants predictor and by-subject intercepts.

4.2.6.3 A-NADL task

We largely followed Lammertink, van Witteloostuijn et al. (2019) in our analysis of the online RT measure of the A-NADL task. Filler trials (20% of trials) and incorrect responses and cases in which no response was given (7.6% of target and non-target trials) were removed prior to analysis.

As in the SRT task, learning during the exposure phase of the A-NADL is assessed as the difference between RTs to predictable input in rule blocks and RTs to pseudo-random input in the disruption block (research question one).

Therefore, in order to find out whether we find evidence of learning during exposure, the linear mixed effects model fitted normalized RTs as a function of the within-participant predictor Block (i.e. rule blocks 3 and 5 versus disruption block 4). Block was orthogonally coded into two contrasts: the effect of learning (i.e. disruption block 4 coded as $-\frac{2}{3}$ vs. rule blocks 3 and 5 coded as $+\frac{1}{3}$ each) and the contrast between the two rule blocks (block 3 vs. block 5 coded as $-\frac{1}{2}$ and $+\frac{1}{2}$ respectively). The second research question, which pertains to the effect of group, was investigated through the interaction between our measure of learning and the between-participant predictor Group (control versus dyslexia). Several other predictors are considered, since these may influence the findings of the model. These included the within-participants predictor Target Type (i.e. target or non-target, coded as $-\frac{1}{2}$ and $+\frac{1}{2}$ respectively) and the between-participants counter-balancing factor Experiment Version (i.e. attending to *lut* coded as $-\frac{1}{2}$ vs. *mip* coded as $+\frac{1}{2}$). Finally, Age was included in the model as an exploratory between-participants predictor. The random effects structure included by-subject and by-item random intercepts and by-subject random slopes for Block and Target Type and by-item random slopes for Experiment Version. The random effect of item refers to the individual *X*-elements used in the familiarization phase of the A-NADL ($N = 24$). By-item random slopes for group and the interaction between experiment version and group were removed. Similarly, by-subject random slopes for the interaction between Block and Target Type were removed. This was done because these resulted in near-perfect correlations, which means that the model was overparameterized. This removal did not result in a decrease in the fit of the model ($\chi^2 = 8.178$, $df = 18$, $p = .98$).

The offline measure of the A-NADL task consisted of yes/no responses to individual items in the GJT, which were coded as 1 (endorsements) or 0 (rejections). The data that served as input to the generalized linear mixed effects model was thus the proportion of endorsements versus rejections (i.e. endorsement rates). No data was removed prior to analysis of the offline GJT.

Importantly, whether an item is endorsed or rejected does not yet inform us about learning, since accuracy depends on the grammaticality of the item (i.e. whether the item adheres to the A-NADL structure or not). To assess whether children showed evidence of learning in the offline measure of the A-NADL (research question one), the model estimated the within-participants effect of Grammaticality (grammatical vs. ungrammatical items orthogonally coded as +

1/2 and - 1/2 respectively) on endorsement rates. The interaction between Grammaticality and Group would provide evidence of a potential difference in performance between children with and without dyslexia (research question two). Since test items either tested familiar *X*-elements or generalization through novel *X*-elements, the within-participants predictor Generalization (no coded as - 1/2 vs. yes coded as + 1/2) was included in the model. Finally, following the online model of the A-NADL, the offline model contained the between-participant counter-balancing predictor Experiment Version and the exploratory between-participant predictor Age. By-subject and by-item random intercepts and by-subject random slopes for Grammaticality and Target Type and random slopes for Experiment Version were included in the random effects. As in the online measure of the A-NADL, the random effect of item refers to the individual *X*-elements used in the test phase of the A-NADL ($N = 16$). By-item random slopes for group and the interaction between experiment version and group and by-subject random slopes for the interaction between Block and Target Type were removed due to overparameterization. Importantly, the fit of the model did not decline ($\chi^2 = 1.793$, $df = 11$, $p = .999$). In order to compute the CIs and *p*-values of the final model, the effect of Grammaticality, which was non-significant, was removed from the random effects structure.

4.3 Results

We focus on confirmatory analyses aimed at answering our research questions. Each time, we separately present results of some of the exploratory analyses, which are not related to our research questions but may nevertheless be interesting (cf. Wagenmakers et al., 2012). Since multiple measures were used to answer our research questions in the VSL and A-NADL tasks, all CIs (and associated significance criteria for *p*-values) of confirmatory results were Bonferroni-corrected to keep the overall false detection rate at 0.05. In the VSL, we used four measures to assess learning (i.e. two online measures: the effect of element and the effect of element in interaction with time, and two offline measures: 2-AFC and 3-AFC accuracy) and thus CIs were corrected for quadruple testing (CIs thereby correspond to a false detection rate of 0.05 / 4 = 0.0125 for each effect, i.e. we have 98.75% CIs). CIs were corrected for double

testing in the A-NADL (i.e. one online and one offline measure), resulting in 97.5% CIs.

As suggested by reviewers, supplementary analyses were conducted including the order of the statistical learning tasks as described in the general procedure as an additional predictor (see OSF for *R* Markdown and html files containing supplementary analyses). Since task order did not interact with our measures of learning (all t and z values < 1.8) and did not result in three-way interactions with our measures of learning and group (all t and z values < 1.8), results from the three testing orders were collapsed in subsequent sections that describe the results of the SRT, VSL and A-NADL tasks.

4.3.1 SRT task

Overall accuracy for both the TD (M = 93.8%) and dyslexia (M = 94.3%) groups was high, indicating that children attended to the task. Figure 4.2 presents the mean normalized RTs to accurate trials across the blocks of the SRT task. RTs were significantly shorter to structured input in sequence blocks 5 and 7 than to random input in disruption block 6 (Δz = -0.276, 95% CI [-0.329 … -0.223], t = -10.292, p = $7.5 \cdot 10^{-9}$), indicating an effect of learning the SRT sequence in children when collapsing over groups. Group did not significantly influence the difference in RT to structured as opposed to unstructured input (Δz = -0.027, 95% CI [-0.133 … +0.079], t = -0.507, p = .61). In other words, there is no evidence for a difference in performance between children with and without dyslexia on the SRT task.

As mentioned, our model provides us with some exploratory findings. Firstly, no significant difference is found between RTs in the two structured blocks (Δz = +0.055, 95% CI [-0.003 … +0.113], t = 1.881, p = .063). The effect of group on the difference in RTs between the two structured blocks also does not reach significance (Δz = +0.073, 95% CI [-0.043 … +0.190], t = 1.248, p = .21). Although participants with dyslexia responded slightly slower than the control group overall, this effect does not reach significance (Δz = +0.102, 95% CI [-0.036 … +0.241], t = 1.462, p = .15). Participants' age was found to influence RTs overall, with shorter RTs with increasing age (Δz = +0.102, 95% CI [-0.286 … +0.1448], t = -6.190, p = $7.5 \cdot 10^{-9}$), but does not interact with the measure of learning and/or with group (see OSF). Lastly, adding attention and verbal short-

term memory to the model does not change the main findings and does not significantly improve the model fit (χ^2 = 5.410, *df* = 12, *p* = .94).

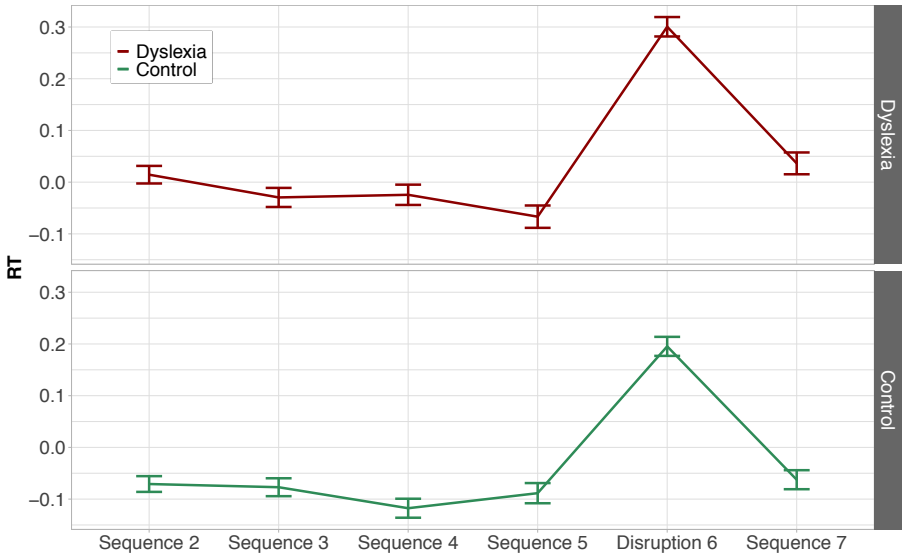


Figure 4.2. SRT task results. Mean normalized RTs (+/- 1 SE) across blocks for participants with dyslexia (top graph; red line) and control participants (bottom graph; green line).

4.3.2 VSL task

4.3.2.1 VSL task: online RT measure

Responses to predictable elements were not significantly shorter as compared to unpredictable elements overall ($\Delta\tilde{\chi}$ = -0.013, 98.75% CI [-0.038 … +0.012], *t* = -1.271, *p* = .21) and there was no evidence of an effect of time in interaction with the online measure of learning (i.e. the difference between predictable and unpredictable elements; $\Delta\tilde{\chi}$ = -0.002, 98.75% CI [-0.024 … +0.019], *t* = -0.276, *p* = .78). Thus, we find no evidence of online sensitivity to the statistical structure in the VSL task. See Figure 4.3 for the mean normalized RTs to predictable and unpredictable elements across repetitions of triplets. The two-way interaction between the measure of learning and group ($\Delta\tilde{\chi}$ = +0.005, 98.75% CI [-0.038 … +0.047], *t* = 0.265 *p* = .79) and three-way interaction including time ($\Delta\tilde{\chi}$ =

+0.024, 98.75% CI [-0.019 … +0.067], t = 1.404, p = .16) were both non-significant. We have no evidence that children with dyslexia perform the online VSL task differently than their TD peers.

The first exploratory finding is that participants with dyslexia responded slightly slower than participants in the control group, but this was not statistically significant ($\Delta\tilde{\chi}$ = 0.009, 95% CI [-0.250 … +0.267], t = 0.066, p = .95). Secondly, RTs were found to be significantly shorter to element 2 than to element 3 within triplets ($\Delta\tilde{\chi}$ = 0.045, 95% CI [+0.019 … +0.070], t = 3.501, p = .00057) and this effect was significantly larger in alien set A than in alien set B ($\Delta\tilde{\chi}$ = -0.190, 95 % CI [-0.254 … -0.126], t = -5.889, $4.6 \cdot 10^{-9}$). Note that there is no significant interaction between the difference in RTs to predictable elements and group (or alien set and group): there is no evidence that children with and without dyslexia perform differently with respect to the difference in RTs to predictable elements 2 and 3 (see OSF). Adding attention and verbal short-term memory to the model does not significantly improve the model fit (χ^2 = 72.296, df = 96, p = .97) or influence the main findings regarding either research question.

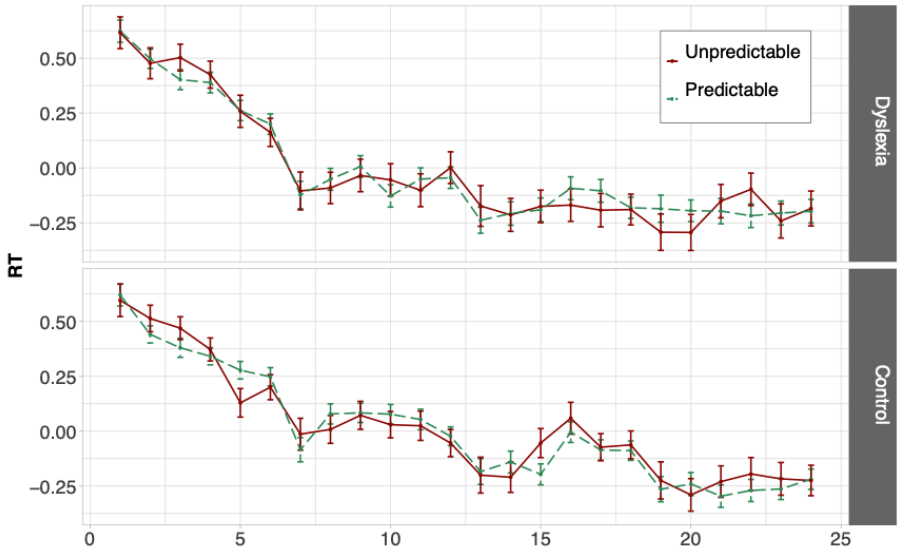


Figure 4.3. VSL task online RT measure results. Mean normalized RTs (+/- 1 SE) to predictable (i.e. elements 2 and 3 within triplets; green dashed lines) and unpredictable (i.e. element 1; red solid lines) elements across repetitions of triplets during the exposure phase for participants with dyslexia (top graph) and control participants (bottom graph).

4.3.2.2 VSL task: offline 3-AFC and 2-AFC measures

Figure 4.4 presents the raw data of the offline test phase of the VSL. In our models, performance was estimated to be 18% and 15% above chance level in the 3-AFC and 2-AFC questions respectively, which was significant in both cases (3-AFC: probability estimate = .520, 98.75% CI = [.462 … .579], $p = 1.8 \cdot 10^{-10}$; 2-AFC: probability estimate = .653, 98.75% CI = [.600 … .704], $p = 3.0 \cdot 10^{-10}$). This means that, collapsing over group, children's offline performance reveals learning in the VSL task. Pertaining to the second aim of our analysis, no significant effect of group was found on performance on 3-AFC (odds ratio estimate = 1.056, 98.75% CI = [0.659 … 1.695], $p = .77$) and 2-AFC (odds ratio estimate = 1.108, 98.75% CI = [0.701 ... 1.751], $p = .57$) questions. Hence, there is no evidence that children with dyslexia perform the offline VSL tasks differently than their TD peers.

The first exploratory finding that should be noted is a significant interaction between alien set and group in the 3-AFC model (odds ratio estimate = 2.699, 95% CI = [1.298 ... 5.662], $p = .0084$): participants with dyslexia performed better in alien set B than in alien set A, and the opposite pattern is observed in the control group. This interaction does not reach significance in the model of 2-AFC performance (odds ratio estimate = 1.839, 95% CI = [0.906 ... 3.766], $p = .091$). Once again, adding attention and verbal short-term memory to the offline models does not significantly improve the model fit for either 3-AFC ($\chi^2 = 18.242$, $df = 16$, $p = .31$) or 2-AFC ($\chi^2 = 21.884$, $df = 16$, $p = .15$) questions and does not change the main findings of either model.

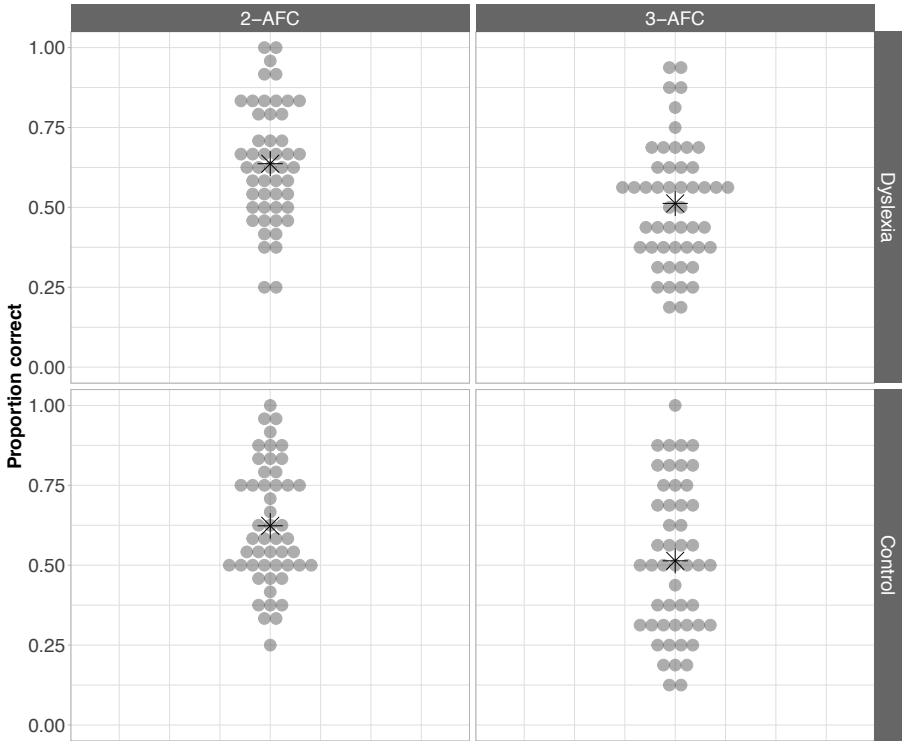


Figure 4.4. VSL offline 3-AFC and 2-AFC measures results. Proportion of correct responses on 2-AFC (left, chance level = .500) and 3-AFC questions (right, chance level = .333) for participants with dyslexia (top) and control participants (bottom). Dots indicate individual scores, while the group mean is indicated using a black asterisk.

4.3.3 A-NADL

4.3.3.1 A-NADL task: online RT measure

Overall accuracy during the online phase of the A-NADL was found to be high for both groups (TD: M = 95.5%, DD: M = 89.2%) indicating that participants attended to the task. Figure 4.5 presents the mean normalized RTs to targets and nontargets across the blocks of the A-NADL experiment. As predicted, RTs in rule blocks were significantly shorter than in the disruption block ($\Delta_{\tilde{\zeta}}$ = -0.159, 97.5% CI [-0.235 … -0.084], t = -4.796, p = $4.9 \cdot 10^{-6}$). Thus, collapsing over group, we find evidence of online sensitivity to the NADL structure. There was no

significant interaction between the effect of learning and group ($\Delta\tilde{\chi}$ = +0.011, 97.5% CI [-0.135 … +0.157], t = 0.167, p = .87). In other words, we find no evidence of a difference in online sensitivity to the A-NADL task between children with and without dyslexia.

Our model also provides us with an exploratory effect of target type, such that RTs were significantly shorter for the stimuli that the participants were attending to than those that were unattended ($\Delta\tilde{\chi}$ = +0.199, 95% CI [+0.156 … +0.242], t = 9.226, p = $2.7 \cdot 10^{-15}$), and experiment version, such that RTs were shorter for participants who attended *lut* than for those that attended *mip* ($\Delta\tilde{\chi}$ = +0.176, 95% CI [+0.018 … +0.335], t = 2.197, p = .030). Additionally, target type and experiment version interact with one another and with our measure of learning (i.e. structured versus disruption blocks) in a three-way interaction ($\Delta\tilde{\chi}$ = +0.168, 95% CI [+0.014 … +0.322], t = 2.134, p = .033). Thus, we find evidence that the effect of learning is enhanced in targets vs. nontargets, especially when children received the version of the A-NADL where they were instructed to attend *lut*. Crucially, the main findings regarding our second research question are not influenced by these exploratory results: we find no significant interactions with the effect of group (see OSF). We found no significant difference in RTs to the two rule blocks included in analyses (i.e. rule block 3 vs. rule block 5; $\Delta\tilde{\chi}$ = +0.056, 95% CI [-0.012 … +0.125], t = 1.624, p = .11) and there was no significant interaction between this difference in RTs and group ($\Delta\tilde{\chi}$ = -0.050, 95% CI [-0.187 … +0.088], t = -0.718, p = .47). There is a marginally significant difference between participants with dyslexia and the TD participants in their overall RTs, with slower responses in the group of participants with dyslexia ($\Delta\tilde{\chi}$ = +0.152, 95% CI [-0.002 … +0.307], t = 1.950, p = .053). adding attention and verbal short-term memory to the model does not significantly improve the model fit (χ^2 = 46.271, df = 48, p = .54) and does not influence the main findings of the online measure of the A-NADL task.

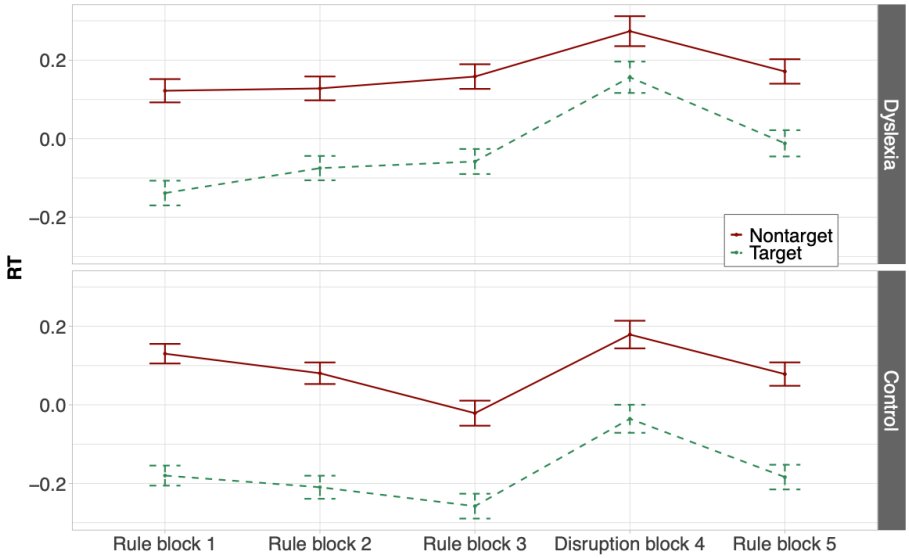


Figure 4.5. A-NADL online RT measure results. Mean normalized RTs (+/- 1 SE) to nontarget (i.e. non-attended *b*-element; red solid line) and target (i.e. attended *b*-element; green dashed line) items across blocks for participants with dyslexia (top graph) and participants in the control group (bottom graph).

4.3.3.2 A-NADL task: offline GJT measure

Figure 4.6 presents the raw proportion of items endorsed (i.e. accepted versus rejected) for participants with and without dyslexia on both grammatical and ungrammatical items in the offline phase of the A-NADL task. The model estimated that the effect of grammaticality on endorsement rates did not reach significance (odds ratio estimate = 1.123, 97.5% Wald CI = [0.592 ... 2.130], p = .68). Hence, we find no evidence of learning in children's offline performance on the A-NADL. As for our second research question, we find no significant interaction between the effect of grammaticality and group (odds ratio estimate = .760, 97.5% Wald CI = [0.421 ... 1.369], p = .30). Therefore, we find no evidence of a difference in performance on the offline measure of the A-NADL between children with and without dyslexia.

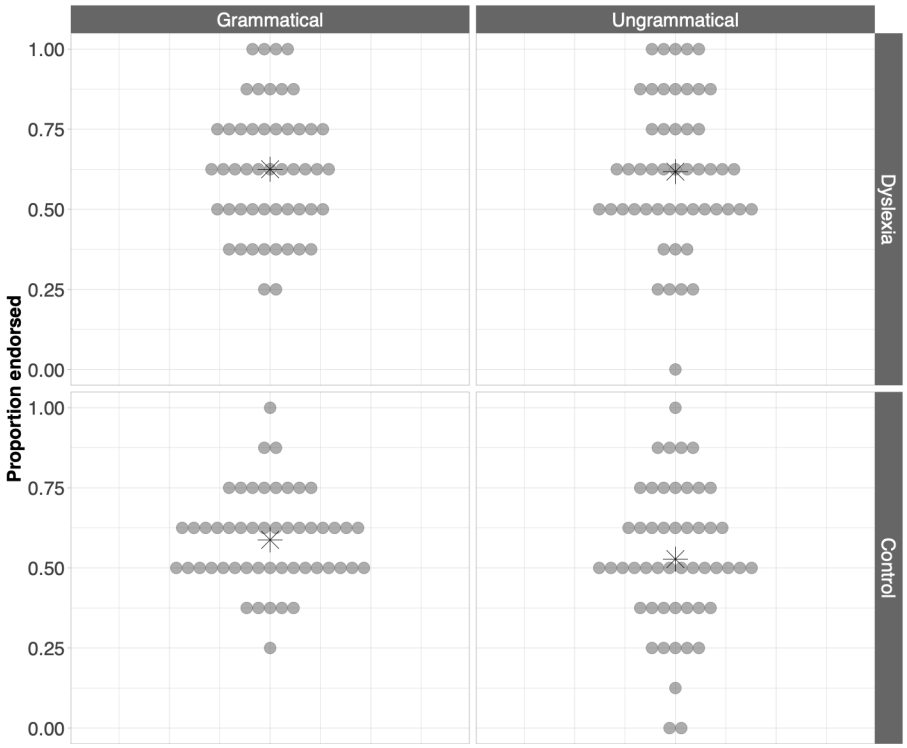


Figure 4.6. A-NADL offline GJT measure results. Proportion of endorsements (i.e. proportion of items endorsed as opposed to rejected; chance level = .500) in the GJT for grammatical (left) and ungrammatical (right) items of participants with dyslexia (top) and control participants (bottom). Dots indicate individual scores, while the group mean is indicated using a black asterisk.

Besides these confirmatory findings, the model revealed a significant yes-bias in the offline measure of the A-NADL (odds estimate of the intercept = 1.507, 95% CI = [1.113 ... 2.050], p = .011): children are more likely to endorse items as opposed to reject them overall. This effect was significantly larger in the group of participants with dyslexia as opposed to the participants in the control group (odds ratio estimate = 1.362, 95% CI = [1.016 ... 1.840], p = .039), reflecting a larger yes-bias in children with dyslexia than in TD children. Secondly, we found that test items that contained a novel *X*-element were endorsed significantly less often than those that contained an *X*-element which had been heard during familiarization (odds ratio estimate = 0.335, 95% CI = [0.183 ... 0.608], p = .0012). Again, this effect is significantly larger in participants

with dyslexia when compared to the TD participants (odds ratio estimate = 1.916, 95% CI = [1.113 ... 3.335], p = .019). We conclude that the endorsement preference (i.e. yes-bias) for familiar over novel *X*-elements is greater for children with dyslexia than for TD children. Note that these effects do not interact with our measures of learning (i.e. grammaticality) or the effect of group and thus do not influence the confirmatory results. As for our previous measures of statistical learning, the findings regarding the offline measure of the A-NADL are not significantly affected by adding sustained attention and short-term memory to the model and the fit of the model is not significantly affected (χ^2 = 44.369, df = 32, p = .072).

4.4 Discussion

The present study investigated statistical learning in children with dyslexia across three different experimental paradigms in a single sample, including the VSL and A-NADL paradigms that had not previously been used in child samples with dyslexia. We aimed to overcome methodological concerns as discussed in the introduction (e.g. Arciuli, 2017; Arciuli & Conway, 2017; West et al., 2017) by assessing learning through a range of online and offline measures and by controlling for group differences in underlying cognitive skills including memory and attention. Across the three statistical learning tasks, we see the same pattern of results: we find evidence of learning when we collapse over groups and we find no evidence of a difference in performance between children with dyslexia and their TD peers. In all analyses, these results remain unchanged after controlling for individual differences in short-term memory and sustained attention. Similarly, the main findings across all statistical learning measures are unaffected by participants' age. Thus, this study finds no evidence in support of (or against) a (domain-general) statistical learning deficit. Thus, our study does not lend support for the hypothesis that a statistical learning deficit is the underlying cause of the literacy problems experienced by children with dyslexia. In the following sections we will elaborate on these findings and their implications.

4.4.1 Measuring statistical learning in child populations

Although overall the same pattern of results arises such that 8- to 11-year-old children show (on- or offline) sensitivity to the statistical structures presented in the SRT, VSL and A-NADL tasks, the measures in the present study differed in their ability to detect learning in this age group. In the VSL task, children learned the structure as indicated by above chance performance on the offline 3-AFC and 2-AFC question, but the online measure did not reveal evidence of learning during the exposure phase. The fact that the children in the present study show offline learning is in line with previous studies that have indicated that, while offline tasks are problematic in younger school-aged children, performance increases between the ages of 5 and 12 (Arciuli & Simpson, 2011; Raviv & Arnon, 2017). Thus, for the VSL paradigm, the offline 3-AFC and 2-AFC questions used here have been demonstrated to be sensitive to learning in children between 8 and 11 years of age. However, we failed to replicate studies that suggested the added value of using the online RT measure of the self-paced VSL in adults and children: participants in these studies responded more slowly to unpredictable stimuli than to predictable stimuli (Siegelman et al., 2018; van Witteloostuijn, Lammertink et al., 2019, see chapter 2). This failure to replicate could be due to small changes in the methodological design but may also be an indication that the RT measure of the VSL is not reliable enough to study performance in child populations. Since the observed difference in RTs between predictable and unpredictable elements across the experiment was deemed small in children between 5 and 8 years of age (van Witteloostuijn, Lammertink et al., 2019, see chapter 2), this effect may be too small to reliably detect across studies and across samples. Future research should further investigate the usefulness of such an online measure and/or alternative online measures when studying statistical learning through the VSL paradigm in children.

The A-NADL task revealed the reversed outcome: children were found to show online sensitivity to the nonadjacent structure, as indicated by an increase in RTs in the disruption block as opposed to RTs in the surrounding rule blocks, while there was no evidence that children endorse more grammatical than ungrammatical items in the offline test phase. This pattern of findings regarding the A-NADL replicates earlier findings in younger TD children by Lammertink, van Witteloostuijn et al. (2019), who report online learning but null findings on

2-AFC questions in 5- to 8-year-olds. As suggested there, the insensitivity of offline tasks could be due to children's difficulties with the meta-linguistic nature of this type of questions. The offline task used for the A-NADL in the present study, the GJT, could contribute to these difficulties, since we have evidence that children are more likely to endorse items than to reject them (i.e. yes-bias). Additionally, we found evidence that children are more likely to endorse items that contain a familiar *X*-element than an unfamiliar *X*-element regardless of their grammaticality. This suggests that children were focused on the *X*-element when answering "yes" or "no". More sensitive offline measures need to be developed to assess the outcome of the learning process in the A-NADL task by children. Importantly, however, the online measure has been shown to be a reliable measure of A-NADL in children, as we replicated the learning effect as reported by Lammertink, van Witteloostuijn et al. (2019). Therefore, future studies investigating A-NADL performance in children could adopt the online measure of learning (in addition to offline measures) to detect sensitivity to nonadjacent structures in speech during exposure.

4.4.2 Statistical learning in dyslexia

The main aim of the present study was to elucidate the extent of the proposed statistical learning difficulties in children with dyslexia. We did not find evidence of group differences on any of the on- or offline measures of the SRT, VSL or A-NADL tasks. Since these tasks assess statistical learning across domains (visuo-motoric, visual and auditory respectively) and across different types of statistical structures (fixed sequence, adjacent and nonadjacent dependencies respectively), we can conclude that we find no support for (or against) a (domain-)general statistical learning deficit in dyslexia.

Of course, a null result is difficult to interpret and can have many possible explanations beside the actual absence of the effect in reality and beside chance. To ascribe meaning to our findings, we have to show that the effects, if they exist at all, are small. Smallness of an effect can be measured by computing its maximal standardized effect size, i.e. by dividing the maximum absolute raw effect size (the greater absolute bound of the confidence interval) by the residual standard deviation of the relevant model. From these post-hoc effect size calculations, we obtain a maximal standardized effect size of $0.160/0.893 = 0.18$

for the online SRT measure, for VSL we get 0.052/0.647 = 0.08, and for the A-NADL we get 0.168/0.815 = 0.21. Therefore, standardized effect sizes on the measures of all three tasks are below or around 0.20 and can therefore be called "small" (Cohen, 1988). One potential cause for the smallness of the effects could be that the selected subjects do not represent the average child with (or without) dyslexia. However, the children with and without dyslexia in the present study were carefully selected according to strict in- and exclusion criteria. The groups were not seen to differ from one another regarding their age, gender, SES and non-verbal reasoning, and the children with dyslexia showed impairments in tasks measuring reading, spelling and lexical retrieval as is characteristic of the disorder. Similarly, the difficulties with verbal short-term memory and sustained attention found in the present study have previously been reported in other samples of children with dyslexia (Bosse et al., 2007; Buchholz & Davies, 2005; Cowan et al., 2017). These are indications that the group of participants with and without dyslexia are representative of the population as a whole. Another potential explanation for the smallness of the effects could be that the statistical learning tasks used are not suitable to assess the underlying construct of statistical learning. However, since we found evidence of learning overall in all three tasks, these paradigms are able to detect learning in children in this age group. Although the methodologies used to investigate statistical learning in child populations should be improved to achieve a full picture of their statistical learning abilities (i.e. the online measure in the VSL and offline measure in the A-NADL), the methodologies of the present study are sensitive enough to potentially detect group differences between participants with and without dyslexia. To summarize, it seems likely that 8- to 11-year-old children with dyslexia do not experience large problems with SL as assessed through these paradigms when compared to age-matched controls. Put more strongly, the results of the present study do not agree with the hypothesis that a domain-general deficit in statistical learning underlies the literacy problems that we see in individuals with dyslexia.

Whereas these results may appear unexpected, other studies have also reported null results (without discussing the effect size) when investigating differences in statistical learning performance between children with and without dyslexia on tasks tapping into statistical learning abilities (SRT e.g. Deroost et al., 2010; Menghini et al., 2010; Staels & Van den Broek, 2017, AGL: e.g. Nigro et al., 2016; Rüsseler et al., 2006). Recently, authors have reached similar inconclusive results regarding the statistical learning deficit hypothesis of dyslexia

in literature reviews and meta-analyses of the SRT and AGL paradigms (Schmalz et al., 2017; van Witteloostuijn et al., 2017, see chapter 3), because these studies underlined the mixed findings (i.e. some studies report significant group effects, while others do not) and established the presence of a publication bias in the field. Of course, methodological differences between studies may (partially) explain the fact that some studies report significant group effects while others do not, especially when the sought-after effect is likely to be small. As also argued by Schmalz et al. (2019) and Elleman et al. (2019), the relationship between performance on statistical learning tasks and literacy skills (and thus dyslexia) may only appear under specific conditions. For example, the type of statistical learning task used (e.g. its statistical structure, its modality), but also the selection of participant groups (e.g. their age, native language, or cultural differences such as differing dyslexia treatments) may influence findings of individual studies. Furthermore, West et al. (2017) question the relationship between statistical learning abilities and dyslexia (and related language learning impairments) based on the poor reliability of the statistical learning tasks used (SRT, Hebb repetition, and contextual cueing) and the lack of correlations between the statistical learning tasks and performance on tasks assessing language and literacy (see also Schmalz et al., 2019).

To conclude, the mixed pattern of findings in the field, and the smallness of the effects found here, suggest that the difference in performance on statistical learning tasks between participants with and without dyslexia may be small and may only be detected under certain experimental conditions (see also Gabay, Schiff, & Vakil, 2012; Henderson & Warmington, 2017, for dissociations between different statistical learning tasks).

4.4.3 Directions for future research

Although the present study detected learning in all three statistical learning paradigms tested (i.e. SRT, VSL, A-NADL), some measures were shown to be less reliable in detecting learning than others (i.e. online learning in the VSL, offline learning in the A-NADL). Future studies that aim to investigate statistical learning performance in children in general, or the relationship between statistical learning and dyslexia more specifically, should aim to develop tasks that are increasingly suitable for assessing statistical learning abilities in child participants.

Additionally, follow-up research using the SRT task could include an explicit offline test phase or consolidation and retention phases in order to gain a complete picture of SRT performance in children with and without dyslexia (Hedenius et al., 2013; Vicari et al., 2005). This adaptation would also allow for a closer comparison with other SL tasks including both on- and offline phases (e.g. VSL and A-NADL).

Since the potential group effect may be small and susceptible to methodological differences between studies, exact replications and large-scale (cross-linguistic and/or cross-cultural) studies are needed to elucidate whether individuals with dyslexia experience (domain-general) difficulties in the area of statistical learning. Future studies could (a priori; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers et al., 2012) choose to conduct Bayesian analyses in order to potentially find support for the null hypothesis that statistical learning abilities in children with and without dyslexia do not differ. As evidence accumulates, existing meta-analyses (Lum et al., 2013; Schmalz et al., 2017; van Witteloostuijn et al., 2017, see chapter 3) could be updated to include recent and future findings to further clarify the clinical relevance of statistical learning in relation to dyslexia and could be extended to further investigate the potential effects of methodological differences between studies (e.g. type of task used, modality tested, and the age or native language of participants).

4.5 Conclusions

This study examined the performance of children with and without dyslexia on three experimental paradigms assessing statistical learning abilities. Across the SRT, VSL and A-NADL paradigms we find that, taken together, children with and without dyslexia are sensitive to the statistical structures presented to them and we find no evidence of a difference in performance between the two groups. Moreover, the group effects reported on in the present study were found to be small. These findings do not support the hypothesis that a domain-general statistical learning deficit results in the literary problems that are observed in individuals with dyslexia. Although future studies are needed to further investigate the direct contribution of statistical learning abilities to literacy acquisition, both in typical and impaired populations, the clinical relevance of statistical learning in relation to dyslexia is likely to be small.

Chapter 5

The contribution of statistical learning to literacy skills*

Abstract

Purpose: Using an individual differences approach in children with and without dyslexia, this study investigated the hypothesized relationship between statistical learning ability and literacy (reading and spelling) skills.

Methods: We examined the clinical relevance of statistical learning (serial reaction time and visual statistical learning tasks) by controlling for potential confounds at the participant-level (e.g. non-verbal reasoning, attention and phonological skills including rapid automatized naming and phonological short-term memory). 100 Dutch-speaking 8- to 11-year-old children with and without dyslexia participated (50 per group).

Results and conclusions: Replicating earlier work, our results demonstrate that phonological skills contribute to individual differences in literacy attainment. No evidence of a relationship between statistical learning and literacy skills is found above and beyond participant-level variables. We propose that the link between statistical learning and literacy attainment, and therefore its clinical relevance, may be small and strongly influenced by methodological differences between studies. Implications for future research are highlighted in the discussion.

* This chapter is a slightly modified version of a manuscript that is currently under review: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (under review at *Dyslexia*). The contribution of individual differences in statistical learning to reading and spelling performance in children with and without dyslexia.

5.1 Introduction

Reading and spelling skills are crucial for academic success and large individual differences in literacy attainment exist, with dyslexia affecting around 3-10% of the population (e.g. Miles, 2004; Siegel, 2006). Learning to read involves the mapping from letters (i.e. graphemes) to sounds (i.e. phonemes), while spelling involves the same mapping in the reversed order. Ideally, the correspondences between graphemes and phonemes are one-to-one. In many orthographies, however, this mapping is complex: graphemes can refer to multiple phonemes and vice versa. For example, the grapheme <c> in English can be expressed either as the phoneme /s/ or /k/ depending on its context (e.g. *cent* versus *can't*). Although children receive explicit instructions regarding some grapheme–phoneme correspondence patterns in school, their ability to implicitly detect statistical regularities, henceforth "statistical learning", has been proposed as an important underlying learning mechanism. This ability is thought to aid the detection of regularities in grapheme–phoneme correspondences when learning to read and spell (e.g. Arciuli, 2017; 2018; Arciuli & Simpson, 2012; Frost et al., 2013; Treiman, 2018). A domain-general learning deficit has been proposed to be the underlying problem in individuals with dyslexia (Nicolson & Fawcett, 2007; 2011); including problems in the area of statistical learning (e.g. Gabay et al., 2015).

One approach to investigating these hypotheses is to correlate performance on independent statistical learning measures with literacy scores. Studies have used a range of statistical learning tasks, including the visual statistical learning (VSL), auditory statistical learning (ASL), and serial reaction time (SRT) tasks. Importantly, these tasks all measure participants' ability to implicitly track statistical regularities from input. Consistent with the above-mentioned proposals, performance on such tasks has been shown to correlate with word and sentence reading in English-speaking adults and children (VSL: Arciuli & Simpson, 2012; ASL: Qi et al., 2019) and with reading Hebrew as a second language in adults (Frost et al., 2013). Replicating Arciuli and Simpson (2012), the relationship between VSL performance and reading accuracy was shown in Norwegian-speaking children (von Koss Torkildsen et al., 2019) and in an additional sample of English-speaking children (Steacy et al. 2019). Findings by Hung et al. (2018) confirmed this relationship using the SRT task in a group

of English-speaking adolescents. A second approach to studying the relationship between statistical learning and literacy attainment is to compare the statistical learning performance of children and adults with dyslexia to typically developing (TD) peers. In line with the hypothesized statistical learning deficit, several studies using a range of statistical learning measures report that individuals with dyslexia perform poorly relative to control groups (e.g. adults: Menghini et al., 2006; Sigurdardottir et al., 2017; children: Gabay et al., 2015; Jiménez-Fernández et al., 2011; Singh et al., 2018).

While these results are promising, other studies challenge the idea of a (strong) relationship between statistical learning and literacy skills. For example, in a large sample of English-speaking TD children (N = 101), no correlations were observed between statistical learning and measures of reading and spelling (West et al., 2017). In a follow-up study, the authors report similar findings: once attention was controlled for, no evidence for a relationship between statistical learning and reading was found (West, Shanks, & Hulme, 2018). Likewise, Schmalz et al. (2019) did not find evidence of the relationship between statistical learning tasks and reading ability in German-speaking adults. They suggest that failures to replicate the correlation between statistical learning and reading are possibly due to the use of different measures of statistical learning, since low correlations between such measures have been previously reported (e.g. Capel, 2018; Misyak & Christiansen, 2012; Schmalz et al., 2018; Siegelman & Frost, 2015). Furthermore, these null findings have led to questions regarding the reliability of statistical learning measures (e.g. Siegelman et al., 2017b; West et al., 2017), especially in use with child participants (Arnon, 2019a; 2019b). A study examining the link between learning on the SRT task and a range of language skills in English-speaking children with and without developmental language disorder (DLD) similarly found no correlation between SRT performance and reading words or pseudo-words in either group (Clark & Lum, 2017). Null findings also exist regarding statistical learning performance in dyslexia: a number of studies did not find evidence for a difference in performance between participants with dyslexia and non-impaired controls (adults: e.g. Kelly et al., 2002; Pothos & Kirk, 2004; children: e.g. Nigro et al., 2016; Staels & Van den Broeck, 2017), even when using a range of different statistical learning measures within the same pool of participants (adults: Rüsseler et al., 2006; children: van Witteloostuijn et al., 2019, see chapter 4). This mixed pattern of findings in the field (i.e. some studies reporting significant group effects and other studies

finding null results) has prompted literature reviews and meta-analyses in the area of statistical learning in dyslexia to determine the overall effect size (Lum et al., 2013; Schmalz et al., 2017; van Witteloostuijn et al., 2017, see chapter 3). Although findings from these meta-analyses suggest that individuals with dyslexia may have problems with statistical learning when collapsing over studies, authors have raised the issue of a publication bias in the field that likely inflates the observed effect size in meta-analyses (Schmalz et al., 2017; van Witteloostuijn et al., 2017, see chapter 3).

Recently, authors have also stressed the need for research combining the two approaches – correlational studies and the investigation of statistical learning in dyslexia – to further elucidate the relationship between statistical learning and literacy skills (Arciuli, 2018; Arciuli & Conway, 2018). To date, several studies have indicated that statistical learning performance relates to reading ability in participants with and without dyslexia (English adults: Gabay et al., 2015; Howard, Howard, Japikse, & Eden, 2006; Icelandic adults: Sigurdardottir et al., 2017; Hebrew children: Vakil, Lowe, & Goldfus, 2015; Swedish children: Hedenius et al. 2013; Dutch children: van der Kleij, Groen, Segers, & Verhoeven, 2019). However, this finding was not replicated by Nigro, Jiménez-Fernández, Simpson and Defior (2015) in a sample of Spanish-speaking children. They argue that statistical learning may play a less prominent role in learning to read a shallow orthography such as Spanish (i.e. a writing system with a relatively transparent grapheme-to-phoneme mapping; but see conflicting results in other more transparent orthographies such as Icelandic, Swedish and Dutch by Sigurdardottir et al., 2017, Hedenius et al., 2017, and van der Kleij et al., 2019, respectively).

To summarize, although theory suggests a link between statistical learning and literacy skills, experimental studies have not been conclusive: while some find evidence supporting the existence of such a relationship, others do not. Several explanations have been proposed for this mixed pattern of findings, including differences in statistical learning tasks used (Schmalz et al., 2019), potential confounds at the participant-level (e.g. attention, West et al., 2018), and low reliability of statistical learning measures (e.g. Arnon, 2019a; 2019b; Siegelman et al., 2017a; West et al., 2017). Interestingly, the majority of studies investigating the relationship between statistical learning and literacy skills have done so through simple correlations, not considering participant-level variables

(i.e. potential confounds) or other known predictors of reading (cf. Qi et al., 2019; von Koss-Torkildsen et al., 2019). Furthermore, studies of individual differences in statistical learning ability in relation to literacy skills have often not reported the reliability of the statistical learning measures used (but see e.g. West et al., 2017; 2018). Moreover, studies to date have largely focused on reading, thereby disregarding spelling despite its theorized link with statistical learning (Treiman, 2018) and despite the spelling difficulties associated with dyslexia (DSM-V, 2013). In relation to dyslexia, it is important to elucidate the clinical relevance of statistical learning in literacy acquisition (see e.g. Plante & Gómez, 2018, for a discussion regarding the clinical relevance of statistical learning for treatment applications with DLD), since mixed findings suggests that the true correlation may only be small (or may be largely mediated by confounding variables).

5.1.1 The present study

In the present study, we further investigate the relationship between statistical learning and literacy in children with and without dyslexia. We hope to do so comprehensively and reliably by looking at both reading and spelling performance and by using two statistical learning measures that have previously been linked to individual differences in literacy skills: the SRT task (e.g. Hedenius et al., 2013; Howard et al., 2006; van der Kleij et al., 2019) and VSL task (e.g. Arciuli & Simpson, 2012; von Koss Torkildsen et al., 2019). Additionally, we aim to account for participant-level characteristics (age, gender, socio-economic status [SES], diagnosis), general cognitive skills (non-verbal reasoning, sustained attention) and other known predictors of literacy outcomes such as rapid automatized naming (RAN), phonological processing and phonological short-term memory (see e.g. de Bree, Wijnen, & Gerrits, 2009; de Jong & van der Leij, 1999; Furnes & Samuelsson, 2010; Swanson & Howell, 2001; van Setten, Hakvoort, van der Leij, Maurits, & Maassen, 2017; see also Snowling & Melby-Lervåg, 2016 for a meta-analysis). Finally, we report the split-half reliability of the statistical learning measures used in the present study as an indication of their internal consistency and reliability (see also e.g. Arnon, 2019a; Siegelman et al., 2017b). Since the statistical learning measures used in the present study were adapted for use with child participants, we hope to find split-half reliability

coefficients that approach the psychometric standard of $r = .80$ (Nunnally & Bernstein, 2004; Steiner, 2003).

Through a regression analysis, we examined the contributions of statistical learning and the aforementioned predictors to reading and spelling performance, including phonological skills[10], in 100 Dutch-speaking school-aged children with and without dyslexia. The current study uses a data set that was previously published (see van Witteloostuijn et al., 2019, see chapter 4). Whereas van Witteloostuijn et al. (2019, see chapter 4) investigated group differences in statistical learning abilities, we here focus on the individual differences in statistical learning and whether these contribute to variability in literacy attainment. Since children were found to be sensitive to the statistical structures presented to them in the SRT and VSL tasks overall (see van Witteloostuijn et al., 2019, see chapter 4, and see §5.3.1), we expect to find meaningful individual differences that may relate to children's literacy performance. The research questions were as follows:

1. Do phonological skills (RAN letters and pictures, nonword repetition [NWR] and digit span forward tasks) contribute to literacy performance?
2. Does statistical learning ability (SRT and VSL tasks) contribute to literacy performance?

And, if so,

3. Are the contributions of phonological skills and/or statistical learning different for
    a. children with and without dyslexia?
    b. reading and spelling?

---

[10] RAN, NWR and digit span forward are grouped together under the label "phonological skills" for ease of reference. The RAN task requires a complex set of skills: e.g. visual recognition, the integration of visual stimuli with stored representations, and the access and retrieval of the associated phonological representations (Norton & Wolf, 2012). NWR involves existing phonological and lexical representations and phonological short-term memory (Rispens & Baker, 2012), while the digit span assesses phonological short-term memory (e.g. Bull, Espy, & Wiebe, 2008).

Please note that the research questions regarding phonological skills concern a "bycatch" of our statistical analysis and are of a purely replicational nature, while the research questions regarding statistical learning are the primary focus of the present study.

5.2 Methods

5.2.1 Participants

Fifty children with a diagnosis of dyslexia (26 girls, 24 boys, age range 8;4 – 11;2, *M* = 9;10) and 50 age-matched control children (24 girls, 26 boys, age range 8;3 – 11;2, *M* = 9;8) in grades three to five participated. Ten additional children with dyslexia and 4 additional control children were tested but turned out not to meet our pre-determined inclusion criteria (dyslexia: norm score of at most 6 [i.e. 10th percentile] on word and pseudo-word reading; control group: norm score of at least 8 [i.e. 25th percentile]). All 100 children that participated in the present study completed each of the tasks as outlined in §5.2.2. Children with dyslexia were recruited through treatment centers and Facebook support groups for parents, while children in the control group were recruited through primary schools. All parents and children consented to participation prior to testing in accordance with the ethical committee of the Faculty of Humanities of the University of Amsterdam. All participants were native speakers of Dutch and none were diagnosed with (additional) developmental disorders as reported by parents (in the case of participants with dyslexia) and teachers (in the case of control participants). Note that the sample reported here is identical to the sample reported by van Witteloostuijn et al. (2019, see chapter 4), since, as previously stated, the current study is a re-analysis of the same sample, but with a different focus. Also, the control group partly overlaps with studies investigating statistical learning and its relationship with language in children with DLD (Lammertink et al., 2019a; 2019b; 2020).

Table 5.1 presents descriptive statistics on a number of participant characteristics (these descriptive statistics overlap with those presented in van Witteloostuijn et al., 2019, see chapter 4). Children with and without dyslexia were found not to differ significantly from one another regarding their age (*t* =

0.839, $p = .40$), SES ($t = 0.173$, $p = .86$) and non-verbal reasoning skills ($t = 0.041$, $p = .97$). Children with dyslexia achieved marginally significantly lower scores than their TD peers on sustained attention ($t = 1.939$, $p = .055$). These participant characteristics are included in our regression analyses as control variables.

Table 5.1. Descriptive statistics of background measures for children with and without dyslexia.

	Dyslexia (*N* = 50)		**Control (*N* = 50)**	
	Raw	Standardized	Raw	Standardized
Female : Male	26 : 24		24 : 26	
Age	9;10 (0;9)	N/A	9;8 (0;10)	N/A
SES[a]	0.2 (1.2)	N/A	0.2 (1.1)	N/A
Nonverbal reasoning[b]	37.2 (6.6)	55.7 (25.0)	37.3 (8.1)	60.1 (28.1)
Sustained attention[c]	7.0 (2.5)	7.4 (3.3)	7.8 (1.8)	9.1 (3.0)

Note: [a] SES was determined on the basis of postal codes through the *Netherlands Institute for Social Research* (NISR), [b] Nonverbal reasoning was assessed through *Raven's Standard Progressive Matrices* (Raven & Raven, 2003). Raw and standardized scores on nonverbal reasoning represent the number of items answered correctly out of 60 and percentile scores (norm = 50) respectively. [c] Sustained attention was measured using the *Score!* subtest of the Dutch *Test of Everyday Attention* (*TEA-Ch*; Schittekatte et al., 2007). Raw and standardized scores on sustained attention represent the number of items answered correctly out of 10 and norm scores (norm = 10) respectively.

5.2.2 Materials

5.2.2.1 Literacy measures

Children's technical reading skills were assessed using two tests: the reading of single Dutch real words (Brus & Voeten, 1972) and pseudo-words (van den Bos et al., 1994). The task was to read (pseudo-)words as quickly and accurately as possible from a set list of words in a set amount of time (one minute for words, two minutes for pseudo-words) and thus the outcome measure was the number of (pseudo-)words read correctly within the time limit. Dutch spelling was measured through a dictation test consisting of 6 blocks containing 15 words (not including verbs) of increasing difficulty both between and within blocks (Braams & de Vos, 2015). Each child completed two blocks depending on their

grade in school (i.e. the spelling score is the number of words spelled correctly out of 30). Every word was presented orally in a context sentence after which the target word was repeated and the child was required to manually write down the word according to Dutch spelling.

5.2.2.2 Phonological skills

Firstly, children were tested on two subtests of RAN: one containing letters and one containing pictures of common objects (van den Bos & Lutje Spelberg, 2007). Children were instructed to name the letters or pictures as quickly and accurately as possible. Secondly, children's phonological processing and short-term memory was assessed through two tasks: a shortened NWR task (NWR-S: le Clercq et al., 2017) and a forward digit span task (Kort et al., 2008). In the NWR-S, children listened to 22 pre-recorded nonwords and had to repeat them as accurately as possible. All nonwords were between three and five syllables long and were either phonologically likely or unlikely according to Dutch phonotactic probabilities. Children's responses were recorded and scored as either correct or incorrect. In the forward digit span task, children had to repeat sequences of digits of increasing length (2–9 digits) in the correct order. Each level of the task contained two items; to advance to the next level the child had to answer at least one out of two items correctly. Testing was halted once a child answered both items within one level incorrectly.

5.2.2.3 Statistical learning

SRT task

The SRT task used in the present study is identical to the one described by van Witteloostuijn et al. (2019, see chapter 4). Participants were exposed to a single visual stimulus that repeatedly appeared in one of four locations (quadrants) on the tablet screen with a 250 milliseconds interval (Nissen & Bullemer, 1987). They were required to respond to the stimulus' location on the screen by pressing one of four corresponding buttons on a game pad controller as quickly and as accurately as possible. Without the participants' knowledge, the SRT task was divided into seven blocks. In blocks 2 through 5 and block 7, the stimulus followed a predetermined sequence of ten locations (4, 2, 3, 1, 2, 4, 3, 1, 4, 3)

which was repeated six times in each block (i.e. 60 trials per block), while the stimulus was presented in random order during 60 trials in the intervening block 6. Block 1 consisted of 20 random trials to accustom participants to the task and is not included in analysis. Learning of the statistical structure in the SRT task is evidenced by longer reaction times (RTs) to random stimuli (block 6) than to structured stimuli in the surrounding sequence blocks (blocks 5 and 7). Individual scores on the SRT task were computed by subtracting the mean normalized RT to structured input (average of blocks 5 and 7) from the mean normalized RT to unstructured input in block 6.

VSL task

The VSL task used in the present study is identical to the one described by van Witteloostuijn et al. (2019, see chapter 4) and was similar in structure and design to previous studies by Arciuli and Simpson (2011; 2012). Twelve visual stimuli (aliens) were presented one by one on a tablet with touch screen. Unbeknownst to the participants, these twelve stimuli repeatedly appeared in the same four groups of three (i.e. triplets; *ABC*, *DEF*, *GHI*, *JKL*). Learning of this triplet structure was originally assessed through three measures: an online RT and two offline accuracy measures. Since no evidence of learning was found through the online RT measure in children with and without dyslexia (van Witteloostuijn et al., 2019, see chapter 4), we focus on the offline measures of learning.

Prior to the experiment, children were informed that aliens stood in line to go home with a space ship and that they would see all of the aliens one by one. They were instructed to pay attention to the aliens and were told that some of the aliens liked one another and stood in line together. The exposure phase of the VSL task contained four separate blocks, each consisting of six occurrences of each triplet. The same triplet never appeared twice in a row and triplet pairs were never repeated (Arciuli & Simpson, 2011; 2012; Turk-Browne et al., 2005). In between blocks, participants received stickers on a diploma. A cover task was inserted in the exposure phase to ensure that children paid attention to the stimulus stream (Arciuli & Simpson, 2011; 2012). Three individual stimuli per block appeared twice in a row and children had to respond to a repeated stimulus by pressing the alien on the screen. Each stimulus within each triplet was repeated once during the exposure phase (e.g. the triplet *ABC* occurs once as *AABC*, *ABBC*, and *ABCC*) and three distinct triplets contained a repetition in random

positions in each of the four blocks, again all three stimulus positions within triplets once (e.g. *AABC*, *DEEF*, *GHII*).

Subsequent to exposure, children were tested on their knowledge of the triplet structure. Using the same set of 12 visual stimuli, four foil triplets (*AEI*, *DHL*, *GKC*, *JBF*) were created for use in the offline test phase that consisted of 40 multiple-choice questions. 16 three-alternative forced-choice (3-AFC) questions in which children had to fill in a missing stimulus (chance level = .333) were followed by 24 2-AFC questions in which they had to pick the more familiar group of aliens (chance level = .500). Both 3-AFC and 2-AFC question blocks were introduced through two practice items during which children were encouraged to make a guess when they were uncertain of the correct response. Individual scores on the VSL task represent the number of items answered correctly out of 16 and 24 on 3-AFC and 2-AFC questions respectively.

5.2.3 General procedure

The SRT and VSL tasks were programmed and run using E-prime 2.0 (Psychology Software Tools, 2012; Schneider et al., 2012) on a Windows Surface 3 tablet with touch screen. Pre-recorded auditory instructions (SRT) and stimuli (NWR-S) were played over Sennheiser HD 201 headphones. Responses in the SRT tasks were given through a Trust wired GXT540 gamepad controller. Children's responses during the reading, RAN and NWR-S tasks were recorded using an Olympus DP-211 voice recorder.

As previously mentioned in §2.1, children were tested in the context of a larger project. An experimenter administered a battery of tasks one-on-one in a quiet room either at the child's home or school. Testing took place in three sessions that lasted around an hour each. The statistical learning tasks were tested in separate sessions along with a number of other measures. The order of the test sessions and the tasks within sessions were counter-balanced: participants were randomly assigned to one out of six testing orders.

5.2.4 Data scoring and analysis

We performed a linear regression analysis through the "lm" function in *R* software to assess the contribution of a number of predictors in explaining individual variation in reading and spelling attainment combined in a single model. 95% confidence intervals (CIs) were computed through the "confint" function and were used to compare the contribution of predictors to reading versus spelling (research question 3b). Predictors in the model included control variables (group membership, age, gender, SES, non-verbal reasoning and sustained attention), phonological skill measures (RAN letters, RAN pictures, NWR-S and digit span forward[11]) and measures of statistical learning (SRT and VSL). Interactions between group and phonological skills and between group and statistical learning measures were investigated (research question 3a). Significance of individual predictors to reading and spelling combined was assessed through the "Manova" function in the *car* package (version 2.1-5; Fox et al., 2012). In order to answer the first two research questions, we conducted model comparisons between the full model and models from which (1) phonological skill measures and (2) statistical learning measures were removed.

All raw scores on continuous measures were centered and scaled using the "scale" function. Categorical predictors were coded into orthogonal contrasts: Gender was coded such that females were marked as -1/2 and males were marked as +1/2; Group membership was coded such that the control group was marked as -1/2 and the dyslexia group was marked as +1/2. Finally, since both reading and VSL were measured through two subtests (reading words and pseudo-words; VSL 3-AFC and 2-AFC), the averages of the centered and scaled

---

[11] Since we were interested in the overall effect of the phonological skill measures, we attempted to summarize these four subtests through maximum likelihood factor analysis. We aimed to reduce all four subtests to one single factor as a minimum of three variables per factor is required (e.g. Tabachnick & Fidell, 2007). However, one factor was deemed insufficient (χ^2 [2] = 13.65, p = .0011). For completion and transparency, we performed identical confirmatory regression analyses using the single score obtained through factor analysis instead of entering the four phonological skill measures individually (see supplementary analyses on the OSF). This alternative approach does not change the main outcomes of the model.

subtests were used in our analyses. Summary level data and *R* Markdown and html files detailing our analyses are available on the corresponding project page on the Open Science Framework (OSF; https://osf.io/dr72a).

The split-half reliability of the statistical learning tasks was computed using Spearman-Brown corrected Pearson correlations (see also Arnon, 2019a; Siegelman et al., 2017a). In the SRT task, the split-half reliability was calculated for each individual as the correlation between the difference in RT between the random stimuli (block 6) and structured stimuli in the surrounding sequence blocks (i.e. blocks 5 and 7) in *even* versus *odd* trials. This difference in RT was obtained from the linear mixed-effects model through the random slopes of the relevant predictor (i.e. the difference in RT between random and sequence). Similarly, the correlation between the accuracy on even and odd trials in the VSL offline test phases (2-AFC and 3-AFC) was used to calculate the split-half reliabilities (i.e. the random slopes of the intercept). We would like to refer the reader to our OSF project page for more detail regarding the calculations of the split-half reliabilities.

5.3 Results

We first provide the descriptive statistics and group comparisons of the outcome measures and predictors included in our linear regression analysis in §5.3.1. The confirmatory analyses aimed at answering our research questions are subsequently presented in §5.3.2. These consist of the linear regression analyses and model comparisons. §5.3.3 presents exploratory analyses and findings, which do not provide answers to our research questions but may nonetheless be of interest (cf. Wagenmakers et al., 2012). Finally, the results regarding the split-half reliabilities of our statistical learning measures are provided in §5.3.4.

Table 5.2. Descriptive statistics (i.e. means, with *SD*s within brackets) on outcome measures, phonological skills and statistical learning: raw and standardized scores per group.

	Dyslexia (*N* = 50)		**Control (*N* = 50)**	
	Raw	Standardized	Raw	Standardized
Reading words[a]	34.1 (11.7)	3.3 (2.1)	66.3 (11.6)	10.5 (2.2)
Reading pseudo-words[a]	22.0 (8.0)	4.4 (1.6)	61.0 (14.4)	11.1 (2.2)
Spelling[b]	8.4 (4.6)	11.8 (13.7)	18.6 (4.7)	49.9 (24.7)
RAN letters[a]	36.1 (10.4)	5.4 (2.7)	27.2 (5.5)	9.6 (3.1)
RAN pictures[a]	53.2 (10.2)	7.7 (2.7)	44.1 (7.3)	10.7 (2.8)
NWR-S[c]	7.3 (2.7)	N/A	9.7 (3.3)	N/A
Digit span forward[a]	7.3 (1.5)	7.7 (2.6)	8.9 (1.5)	10.7 (2.9)
SRT[c]	0.29 (0.28)	N/A	0.27 (0.27)	N/A
VSL 3-AFC[cd]	8.2 (3.1)	N/A	8.2 (3.8)	N/A
VSL 2-AFC[cd]	15.3 (4.4)	N/A	15.0 (4.5)	N/A

Note: Raw scores: reading words and pseudo-words = the number of words read within the time limit, spelling = the number of words spelled correctly out of 30, RAN = the number of seconds spent on the task (i.e. higher score = lower performance), NWR = the number of nonwords repeated correctly out of 22, Digit span forward = the number of items answered correctly out of 16, SRT = difference in normalized RTs (RT random – RT sequence), VSL = number of items answered correctly out of 16 (3-AFC) and 24 (2-AFC). Standardized scores represent either [a] norm scores (norm = 10) or [b] percentile scores (norm = 50). [c] No standardized scores are present for the NWR-S, SRT and VSL tasks. [d] Chance level on VSL 3-AFC = $^1/_3$ (5.3 items correct out of 16); 2-AFC $^1/_2$ (12 items correct out of 24).

5.3.1 Descriptive statistics and group comparisons

Table 5.2 contains the descriptive statistics of reading, spelling, phonological skills and statistical learning (note once again that these descriptive statistics overlap with those presented in van Witteloostuijn et al., 2019, see chapter 4). As expected, children with dyslexia performed significantly worse than children in the control group on reading (words: $t = 13.83$, $p = 9.1 \cdot 10^{-25}$; pseudo-words: $t = 16.75$, $p = 1.6 \cdot 10^{-30}$), spelling ($t = 11.05$, $p = 9.4 \cdot 10^{-20}$) and phonological skills (RAN letters: $t = 5.421$, $p = 4.3 \cdot 10^{-7}$; RAN pictures: $t = 4.985$, $p = 2.7 \cdot 10^{-6}$; NWR-S: $t = 3.962$, $p = .00014$; digit span forward: $t = 5.36$, $p = 5.5 \cdot 10^{-7}$). Children

learned the statistical structures in the SRT and VSL tasks overall and no evidence of a difference in performance between children with and without dyslexia was found on the statistical learning measures (SRT: $\mathcal{L}_{\tilde{\chi}}$ = -0.027, p = .61; VSL 3-AFC odds ratio estimate = 1.001, p = .996; VSL 2-AFC: odds ratio estimate = 1.076, p = .68; see van Witteloostuijn et al., 2019, see chapter 4). For more detail on the analysis of the statistical learning measures, please see the OSF project page.

5.3.2 Regression model

The outcomes of the full model are presented in Table 5.3 for reading and spelling separately. Our first research question pertained to the contribution of phonological skills to literacy skills overall. Also, we were interested to see whether its contribution differed between children with and without dyslexia (research question 3a). From the "Manova" function results, as presented in Table 5.4, we see that only RAN letters is a significant contributor to literacy outcomes combined, over and above the other predictors in the model (Wilk's λ= .74, F[2,80] = 13.840, p = $6.9 \cdot 10^{-6}$). Additionally, the interaction with Group is significant (Wilk's λ= .90, F[2,80] = 4.600, p = .013), such that the effect of RAN letters is larger for children with dyslexia than for TD children. The interaction between RAN pictures and Group is significant in the same direction (Wilk's λ = .92, F[2,80] = 3.453, p = .036). Regarding differences between reading and spelling (research question 3b), the effect of RAN letters is significantly larger on reading than on spelling since the 95% CIs do not overlap (β = -.28, 95% CI [-.42 ... -.15], p = $6.5 \cdot 10^{-5}$ and β = .096, 95% CI [-.13 ... .33], p = .41 respectively). Importantly, when we compare the full model to a reduced model where the phonological skill measures are removed (RAN letters, RAN pictures, NWR-S and digit span forward), we find that this removal results in a significant decrease in model fit (F[16,162] = 3.771, p = $6.4 \cdot 10^{-6}$). Thus, taken together, the phonological skill measures used in the present study (RAN letters and pictures, NWR-S and digit span forward) contribute to children's literacy performance.

Secondly, and most importantly, we were interested in the contribution of statistical learning to individual differences in literacy attainment (research question 2) and potential differences between children with and without dyslexia (research question 3a). We see that the main effects of SRT and VSL are non-

significant overall (Wilk's λ = .97, F[2,80] = 1.164, p = .32 and Wilk's λ = .99, F[2,80] = 0.496, p = .61 respectively). The interaction between Group and SRT performance approaches significance (Wilk's λ = .94, F[2,80] = 2.330, p = .10). Although the interaction between Group and SRT performance is significant for spelling (β = .29, 95% CI [.02 ... .57], p = .036) but not for reading (β = .13, 95% CI [-.03 ... .29], p = .11), we cannot infer a difference between reading and spelling due to overlapping 95% CIs (research question 3b). Comparing the full model to a reduced model (where SRT and VSL are removed) does not show a significant effect of the removal of statistical learning measures on the model fit (F[8, 162] = 1.134, p = .34). In other words, there is no evidence that the SRT and VSL measures together contribute to literacy performance in children with and without dyslexia (above and beyond our control variables and phonological skill measures).

Table 5.3. Full linear regression model: reading and spelling outcomes separately (*lm*), part 1.

	Reading				**Spelling**			
	β	95% CI	*t*-value	*p*	*β*	95% CI	*t*-value	*p*
Control								
Age	.092	[.006 ... .18]	2.13	.036*	.014	[-.14 ... .16]	.18	.86
Gender	-.0011	[-.17 ... .17]	-.012	.99	-.078	[-.37 ... .22]	-.53	.60
SES	.018	[-.07 ... .10]	.44	.66	.018	[-.13 ... .16]	.25	.80
Raven	.017	[-.09 ... .12]	.32	.75	.20	[.03 ... .38]	2.31	.024*
Attention	.034	[-.05 ... .12]	.80	.42	-.020	[-.16 ... .13]	-.27	.79
Group	-1.23	[-1.45 ... -1.01]	-11.13	< .001*	-1.48	[-1.86 ... -1.10]	-7.77	<.001*
Phonology								
RAN let	-.28	[-.42 ... -.15]	-4.21	< .001*	.096	[-.13 ... .33]	.83	.41
RAN pic	-.10	[-.21 ... .007]	-1.86	.066†	.016	[-.17 ... .21]	.17	.87
NWR-S	.097	[-.004 ... .20]	1.91	.060†	.072	[-.10 ... .25]	.82	.41
DSF	.0052	[-.10 ... .11]	.10	.92	.078	[-.10 ... .25]	.88	.38

Note: RAN let = RAN letters, RAN pic = RAN pictures, DSF = Digit Span Forward. Significant findings ($p \leq .05$) are indicated using an asterisk (*), while trends ($.05 \leq p \leq .10$) are indicated using a cross (†).

Table 5.3. Full linear regression model: reading and spelling outcomes separately (*lm*), part 2.

	Reading				**Spelling**			
	β	95% CI	*t*-value	*p*	β	95% CI	*t*-value	*p*
SL								
SRT	.028	[-.06 ... .11]	.67	.51	.095	[-.05 ... .24]	1.32	.19
VSL	.031	[-.07 ... .13]	.63	.53	-.017	[-.18 ... .15]	-.21	.84
Interactions								
Group*RAN let	.27	[.002 ... .54]	2.00	.048*	-.14	[-.61 ... .33]	-.59	.55
Group*RAN pic	.20	[-.01 ... .42]	1.86	.066†	-.064	[-.44 ... .31]	-.34	.73
Group*NWR-S	-.067	[-.26 ... .13]	-.67	.50	-.13	[-.46 ... .21]	-.73	.47
Group*DSF	.10	[-.10 ... .30]	.99	.33	-.025	[-.37 ... .32]	-.14	.89
Group*SRT	.13	[-.03 ... .29]	1.64	.11	.29	[.02 ... .57]	2.13	.036*
Group*VSL	.074	[-.12 ... .26]	.77	.44	.18	[-.15 ... .51]	1.09	.28

Note: RAN let = RAN letters, RAN pic = RAN pictures, DSF = Digit Span Forward. Significant findings ($p \leq .05$) are indicated using an asterisk (*), while trends ($.05 \leq p \leq .10$) are indicated using a cross (†).

Table 5.4. Full linear regression model: outcomes for reading and spelling combined ("Manova" function)

	Pillai's Trace	***F*(2, 80)**	***p***
Control			
Age	.925	3.24	.044*
Gender	.995	.21	.81
SES	.998	.09	.91
Raven	.918	3.57	.033*
Attention	.981	.78	.46
Group	.320	84.88	< .001*
Phonology			
RAN let	.743	13.84	< .001*
RAN pic	.968	1.30	.28
NWR-S	.949	2.16	.12
DSF	.987	.52	.60
SL			
SRT	.971	1.16	.32
VSL	.988	.50	.61
Interactions			
Group*RAN let	.897	4.60	.013*
Group*RAN pic	.921	3.45	.036*
Group*NWR-S	.992	.31	.74
Group*DSF	.977	.93	.40
Group*SRT	.945	2.33	.10†
Group*VSL	.985	.60	.55

Note: RAN let = RAN letters, RAN pic = RAN pictures, DSF = Digit Span Forward. Significant findings ($p \leq .05$) are indicated using an asterisk (*), while trends ($.05 \leq p \leq .10$) are indicated using a cross (†).

5.3.3 Exploratory results

5.3.3.1 Control variables

As expected, Group is a significant predictor for both literacy measures combined (Wilk's $\lambda = .32$, $F[2,80] = 84.876$, $p = 3.4 \cdot 10^{-20}$), such that children with dyslexia achieve lower scores than their TD peers. Similarly, children's age is found to be a significant predictor of literacy performance (Wilk's $\lambda = .93$,

F[2,80] = 3.237, p = .044). This effect is driven by a significant effect of age on reading (β = .092, 95% CI [.01 ... .18], t = 2.130, p = .036), but not spelling (β = .014, 95% CI [-.14 ... .16], t = .181 p = .86). This is to be expected, since the spelling test used is adapted to children's grade, whereas the reading test is not. The opposite pattern is observed for non-verbal reasoning, which is a significant predictor for spelling (β = 0.20, 95% CI [.03 ... .38], t = 2.308, p = .024) but not reading (β = .017, 95% CI [-.09 ... .12], t = .324, p = .75). Again, the overall effect of non-verbal reasoning on literacy skills combined is found to be significant (Wilk's λ = .92, F[2,80] = 3.566, p = .033).

5.3.3.2 Phonological skills

In the full model, only significant effects are found concerning the RAN letters subtest, suggesting that the effect of the phonological skill measures may be carried largely by RAN letters. Therefore, an exploratory analysis was performed to see whether removing the other measures of phonological skills (i.e. RAN pictures, NWR-S and digit span forward) would result in a decrease in fit of the model. Results reveal that this is not the case, as the model comparison is not significant (F[12, 162] = 1.559, p = .11). This means that there is no evidence that, taken together, RAN pictures, NWR-S and digit span forward contribute to literacy performance above and beyond RAN letters.

5.3.3.3 Statistical learning

Perhaps unexpectedly, we find no evidence that children's VSL performance contributes to literacy scores above and beyond the SRT. To investigate whether the VSL may be of value when the SRT is not considered, we ran an identical model with the SRT measure removed. However, the effects of the VSL remain non-significant both in reading (β = .027, t = .555, p = .58) and in spelling (β = -.024, t = -.279, p = .78) and no interactions between VSL and group are found for either outcome measure (β = .077, t = .802, p = .43 and β = .19, t = 1.099, p = .27 respectively).

We also want to explore the interaction between Group and SRT, which approached significance for reading and spelling combined. For further

investigation, we computed Pearson's correlations between SRT and our literacy outcomes (see *R* markdown and html files for plots). The correlation with spelling was found to be non-significant in the control group (r = -.103, p = .48), whereas it reached significance in the group of children with dyslexia (r = .372, p = .0078). Similar results are observed regarding reading (control group: r = -.229, p = .11, dyslexia group: r = .348, p = .013).

5.3.4 Split-half reliability of the statistical learning measures

As explained in §5.2.4, split-half reliabilities were calculated as a measure of the internal consistency and reliability of the statistical learning measures used in the present study. The split-half reliability for the online measure of learning in the SRT task was found to be r = .71, 95% CI [.58, .81]. For the offline measures of learning in the VSL task, the split-half reliabilities were r = .70, 95% CI [.55 ... .80] and r = .78, 95% CI [.67 ... .85] for 2-AFC and 3-AFC questions respectively. Thus, the split-half reliabilities found for the SRT and VSL tasks used in the present study approach the psychometric standard of r = .80 (see e.g. Nunnally & Bernstein, 1994; Steiner, 2003).

5.4 Discussion

The current study examined the contribution of phonological skills and statistical learning ability to individual differences in reading and spelling performance in children with and without dyslexia. We aimed to do so whilst controlling for potential participant-level confounds including a range of cognitive and phonological skills. This was done in order to investigate whether statistical learning contributes to reading and spelling above and beyond other potential predictors of literacy performance. Our finding that phonological skill measures contribute to literacy scores replicates earlier work (e.g. de Jong & van der Leij, 1999; Swanson & Howell, 2001; Snowling & Melby-Lervåg, 2016). Moreover, its contribution appears to be carried mostly by measures of RAN and the effect is found to be larger for children with dyslexia than for control participants. Regarding the relationship with statistical learning, exploratory simple correlations suggest a (weak) association between SRT performance and literacy

skills exist in participants with dyslexia. No support for (or against) such a link is observed in the control group, or with the VSL task. However, after controlling for the aforementioned participant-level variables, we find no evidence that statistical learning (SRT and VSL) ability contributes to reading and spelling.[12] Regression analysis did not reveal significant differences regarding this relationship between groups (dyslexia versus control) or outcome measures (reading versus spelling). Thus, our results are in agreement with other studies that do not provide evidence for the relationship between statistical learning and literacy skills (e.g. Nigro et al. 2015; West et al. 2017; 2018; Schmalz et al. 2019), despite theoretical claims and experimental evidence of the existence of this relationship from other studies (e.g. Arciuli, 2018; Arciuli & Simpson, 2012; Treiman, 2018; von Koss Torkildsen et al., 2019). Furthermore, our findings highlight the importance of controlling for participant-level variables when investigating the link between SL and literacy attainment.

The absence of evidence for a (strong) relationship between statistical learning and literacy skills in the present study may have a number of explanations. Although these null results may simply be due to chance, several methodological choices may have influenced the outcomes of the present study. Specifically, the statistical learning tasks reported here are not exact replications of those employed in previous studies, and we consider a unique range of participant-level variables. Although the VSL task is identical in statistical structure to the task used in previous studies (Arciuli & Simpson, 2012; Qi et al., 2019; von Koss Torkildsen et al., 2019), it involves a different set of stimuli and a novel online measure of learning during exposure (i.e. the task was self-paced; see Siegelman et al. 2018; van Witteloostuijn, Lammertink et al., 2019, see chapter 2). Similarly, the SRT task resembles tasks used by Hung et al. (2018) and van der Kleij et al. (2019), but notable differences include the sequence to be learned (e.g.

[12] Statistical learning may play a more prominent role in pseudo-word reading than in real word reading, due to the fact that pseudo-words have not been encountered before and therefore readers have to read indirectly through grapheme–phoneme mappings (see e.g. van der Kleij et al., 2019). Thus, we performed identical confirmatory regression analyses using pseudo-word reading as an outcome measure instead of both reading measures combined (see supplementary analyses on the OSF). This alternative analysis provides a similar pattern of results. Most importantly, removing the statistical learning measures from the model does not significantly decrease the model's fit ($F[8, 162] = 1.244$, $p = .28$).

Hung et al., 2018: a 12-item sequence; here: a 10-item sequence), the number of exposures (e.g. van der Kleij et al., 2019: 70 exposures to the sequence prior to the random block; here: 24 exposures prior to the random block), and the visual set-up of the task (e.g. van der Kleij et al., 2019: three locations on the screen presented horizontally; here: four locations presented as a quadrant). Following Schmalz et al. (2019) and Elleman et al. (2019), we suggest that the mixed pattern of findings in the literature examining the association between statistical learning and literacy skills is likely (at least partially) due to such methodological choices: since the true association may be relatively small (or, in fact, zero), it may only appear under certain experimental conditions. These choices could involve the type of statistical structure tested (e.g. adjacent versus nonadjacent dependencies), the modality of the task (e.g. visual versus auditory), the type of task used (e.g. VSL versus SRT), and the type of instruction given to participants (i.e. more or less implicit). Furthermore, current statistical learning tasks are known to show low correlations amongst each other, which may help explain mixed results when investigating the relationship between statistical learning and other cognitive or linguistic skills (e.g. Schmalz et al., 2019; Siegelman & Frost, 2015).

Another explanation previously put forward is the idea that statistical learning may play a less prominent role in more transparent orthographies (such as Dutch, as examined here) than English, since grapheme–phoneme correspondences in these orthographies are potentially easier to acquire through explicit instruction (see e.g. Elleman et al., 2019; Nigro et al., 2015; Schmalz et al., 2019). This seems a less likely explanation, since other studies involving (semi-)transparent orthographies such as Norwegian (von Koss Torkildsen et al., 2019) and Icelandic (Sigurdardottir et al., 2017) reported significant associations between (visual) statistical learning tasks and reading performance, even after considering a range of reading-related abilities (von Koss Torkildsen et al. 2019). Moreover, von Koss Torkildsen et al. (2019) report a comparable effect size as found for English (Arciuli & Simpson, 2012), which suggests similar influences of statistical learning on reading performance in (semi-)transparent and opaque orthographies.

Recently, concerns have been raised about the reliability of statistical learning measures (e.g. Kidd et al., 2017; Siegelman et al., 2015; 2017a; West et al., 2017; 2018), especially in child participants (Arnon, 2019a; 2019b), which limits their appropriateness for studies of individual differences. The statistical

learning measures in the present study had split-half reliabilities of $r = .71$ (SRT) and $r = .70$ and $r = .78$ (VSL 2-AFC and 3-AFC respectively). Previous reports on the reliability of statistical learning measures in children have been less promising, with split-half reliabilities between $r = -.04$ (ASL) and $r = .46$ and $r = .59$ on a VSL (Arnon, 2019a). In their study of the SRT task, West et al. (2017) report split-half reliabilities of between $r = .17$ and $r = .75$. Ideally, the reliability coefficients of psychological measurements reach the value of $r = .80$ (see e.g. Nunnally & Bernstein, 1994; Steiner, 2003). We could say, therefore, that the reliability coefficients of the statistical learning measures used in the present study approach psychometric standards, although it is important to emphasize that there remains room for improvement.

In order to clarify the true relationship between statistical learning and literacy acquisition, an important aim for future research is to develop statistical learning tasks that can be considered reliable, not only regarding split-half reliabilities but also test-retest reliabilities, and are therefore suitable for examining individual differences (see also e.g. Kidd et al., 2017; Siegelman et al., 2017a; 2017b). The development of reliable statistical learning measures that are suitable for use with child participants is especially important (e.g. Arnon, 2019a; 2019b). Additionally, the present state of the field stresses the need for (exact) replications and large-scale (cross-linguistic) studies, preferably using a fixed set of tasks. We would also like to stress the added value of pre-registration and registered reports, which could help minimize problems such as a publication bias in the field and may thereby clarify the nature of the relationship between statistical learning and literacy skills (see also e.g. Schmalz et al., 2017; van Witteloostuijn et al., 2017). Theoretical and pedagogical models of reading and spelling should be extended to incorporate statistical learning in order to enable the formulation of more specific and testable hypotheses for future studies such as "at what stage of learning to read and spell is statistical learning of importance?" and "what type of statistical learning is most closely associated with literacy acquisition?". With the accumulation of evidence, meta-analytic analyses may provide insight into the strength of the relationship between statistical learning and literacy skills, which in turn can clarify its relevance for clinical practice and potential use in treatment for individuals with dyslexia. Meta-regression techniques could inform us about potential moderators of the effect such as participant characteristics (e.g. age, native orthography) and

methodological choices regarding the statistical learning task (e.g. type of structure, modality).

To conclude, the results of the present study fit with the pattern of mixed findings in the field more generally: although we find evidence of correlations between SRT performance and reading and spelling in children with dyslexia (although weak and uncontrolled for potential participant-level confounds), no evidence for a relationship between statistical learning and literacy attainment is found once we consider participant-level characteristics such as age, non-verbal reasoning, attention and phonological skills and when we consider the whole sample of children with and without dyslexia. Although these null results may simply be due to chance, it may also suggest that the link between statistical learning and literacy skills may be less strong than previously hypothesized and is likely influenced by methodological choices made in individual studies.

Chapter 6

Individual differences in statistical learning and grammar*

Abstract

Purpose: Several studies have signalled grammatical difficulties in individuals with developmental dyslexia. These difficulties may stem from a phonological deficit (e.g. Shankweiler et al. 1995), but may alternatively be explained through a domain-general deficit in statistical learning (e.g. Nicolson & Fawcett, 2007). This study investigates grammar in children with and without dyslexia, and whether phonological memory and/or statistical learning ability contribute to individual differences in grammatical performance.

Methods: We administered the standardized Clinical Evaluation of Language Fundamentals (CELF-IV-NL; Kort et al., 2008) "word structure" and "recalling sentences" subtests and measures of phonological memory (digit span, nonword repetition) and statistical learning (serial reaction time, nonadjacent dependency learning) among 8- to 11-year-old children with and without dyslexia (N = 50 per group).

Results and conclusions: Consistent with previous findings, our results show difficulties in the area of grammar, since children with dyslexia achieved lower scores on CELF (word structure: p = .0027, recalling sentences: p = .053). While the two phonological memory measures were found to contribute to individual differences in grammatical performance, no evidence for a relationship with statistical learning was found. An error analysis revealed errors on irregular morphology (e.g. plural and past tense), suggesting problems with lexical retrieval. These findings are discussed in light of theoretical accounts of the underlying deficit in dyslexia.

---

* This chapter is a slightly modified version of a submitted manuscript: van Witteloostuijn, M.T.G., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (submitted to *Applied Psycholinguistics*). Grammatical difficulties in children with dyslexia: The contributions of individual differences in phonological memory and statistical learning.

6.1 Introduction

Developmental dyslexia (henceforth "dyslexia") is a learning disability that is characterized by impaired reading and spelling despite normal intelligence and educational opportunities, and in absence of sensory impairments (DSM-V, 2013; Snowling, 2001). Individuals with (a familial risk of) dyslexia are known to experience difficulties in the area of phonological skills (Vellutino, Fletcher, Snowling, & Scanlon, 2004; see Melby-Lervåg et al., 2012 for a meta-analysis), which has led to the predominant view that the literacy impairments in dyslexia stem from an underlying phonological deficit. When learning to read and spell, children must acquire the correspondences between letters and sounds (i.e. graphemes and phonemes). If, however, the processing, storage and/or representation of phonological information is impaired, children experience difficulties in the acquisition of grapheme–phoneme mappings that in turn result in problems with literacy acquisition (e.g. Ramus & Szenkovits, 2008).

Over time, researchers have uncovered cognitive impairments in individuals with (a familial risk of) dyslexia in addition to literacy and phonological skills. These (may) involve general skills such as the processing of visual and auditory input (e.g. Stein & Walsh, 1997; Tallal, 2004), attention (e.g. Facoetti et al., 2000), and motor functioning (e.g. Ramus, 2003), but may also extend to language domains other than literacy and phonology, such as inflectional morphology and syntax (e.g. Rispens & Been, 2007; Robertson & Joanisse, 2010; Scarborough 1990). Together, these observations have led to suggestions of a more general learning deficit in dyslexia (Nicolson & Fawcett, 2007; 2011), i.e. a deficit of the domain-general ability to detect statistical patterns in sensory input, including spoken and written language (henceforth "statistical learning"; Gabay et al., 2015; Lum et al., 2013).

The aim of the present study is twofold: (1) to investigate the performance of children with and without dyslexia on measures assessing inflectional morphology and syntax, and, most importantly, (2) to examine whether children's performance in these domains can be explained by individual differences in phonological processing and memory and/or statistical learning ability. In doing so, we contribute to the existing literature on grammatical ability in children with dyslexia and enhance our understanding of their difficulties in this area. Most importantly, we hope to provide novel insights into the underlying

cause of the linguistic difficulties observed in dyslexia by investigating two opposing theories (i.e. phonological or statistical learning deficit).

6.1.1 Grammatical difficulties in children with dyslexia

Besides pronounced deficiencies in the areas of literacy skills and phonology, individuals with (a familial risk of) dyslexia have been shown to experience delays in oral language development in early childhood (see Snowling & Melby-Lervåg, 2016, for a meta-analysis). Studies of spoken language skills in young children with a familial risk of dyslexia have shown that they produce shorter sentences of lower syntactic complexity and achieve lower vocabulary scores than typically developing (TD) children (e.g. van Alphen, de Bree, Gerrits, de Jong, Wilsenach, & Wijnen, 2004; Chen, Wijnen, Koster, & Schnack, 2017; Koster, Been, Krikhaar, Zwarts, Diepstra, & van Leeuwen, 2005; Lyytinen et al., 2001; Ramus, Marshall, Rosen, & van der Lely, 2013; Scarborough, 1990; Snowling & Melby-Lervåg, 2016). Furthermore, when school-aged children with (a familial risk of) dyslexia are compared to their TD peers, they are found to achieve lower scores on standardized tests of grammar (e.g. the *Clinical Evaluation of Language Fundamentals* [CELF]; total language score: McArthur et al., 2000; word structure subtest: Joanisse et al., 2000; recalling sentences subtest: Ramus et al., 2013; Finnish inflectional morphology test: Aro, Eklund, Nurmi, & Poikkeus, 2012; Lyytinen et al., 2001, but see e.g. Carroll & Myers, 2010). Similarly, there are indications that (pre)school-aged children with (a familial risk of) dyslexia perform more poorly on experimental tasks that assess inflectional morphology, including pluralization (Joanisse et al., 2000), subject-verb agreement (Jiménez et al., 2004; Nash, Hulme, Gooch, & Snowling, 2013; Rispens et al., 2004; Rispens & Been, 2007) and past tense formation (de Bree & Kerkhoff, 2010; Joanisse et al., 2000; Nash et al., 2013; but see Rispens, de Bree, & Kerkhoff, 2014). The same holds for the comprehension of sentences (Robertson & Joanisse, 2010) and the correct interpretation (and production) of complex syntactical structures such as passive sentences (Reggiani, 2010; Shankweiler et al., 1995), relative clauses (Bar-Shalom, Crain, & Shankweiler, 1993; Mann et al., 1984; Shankweiler et al., 1995), and referential pronouns (Waltzman & Cairns, 2000).

Note, however, that null findings regarding the group comparisons with age-matched TD children have also been reported in the literature: on

standardized tests of grammar (Carroll & Myers, 2010; Ramus et al., 2013) and on experimental tasks examining inflectional morphology and syntax (e.g. Rispens et al., 2014; Ramus et al., 2013). It is also noteworthy that most studies of grammar in dyslexia have focused on pre-school and early-school-aged children (i.e. up until the age of 8; but see Ramus et al., 2013; Rispens et al., 2014; Robertson & Joanisse, 2010). Thus, the reported difficulties in the area of inflectional morphology and syntax in individuals with dyslexia may be restricted to specific grammatical processes (or may be described as subtle; see Rispens et al., 2014), and not much is known about the persistence of these difficulties into later childhood.

The abovementioned oral language difficulties in children with dyslexia are reminiscent of developmental language disorder (DLD; previously known as specific language impairment [SLI]; Bishop, Snowling, Thompson, & Greenhalgh, 2017), a disorder that is defined by oral language problems and pronounced difficulties in the areas of morphology and syntax. Dyslexia and DLD are distinct diagnoses that can co-occur within a single child. The behavioral overlap between the two disorders is known to be high (e.g. McArthur et al., 2000; Catts et al., 2005), which has raised the question whether the two disorders should be viewed as distinct or as two points on a single continuum (e.g. Bishop & Snowling, 2004). Although we are aware of this line of research, we here focus on children with a diagnosis of dyslexia and investigate the nature and extent of their grammatical difficulties.

6.1.2 Theories of dyslexia: phonological deficit and statistical learning deficit

Theories of the underlying cause of dyslexia should not only account for the impairments in the area of reading and spelling, but should also be able to explain the abovementioned difficulties with inflectional morphology and syntax. In line with the dominant view that dyslexia originates from a deficit in phonological skills, grammatical problems in dyslexia have been theorized to be "further symptoms of an underlying phonological weakness" (Shankweiler et al., 1995, p. 149). This idea is supported by evidence that children with dyslexia are especially impaired in morpho-phonology – morphological processes that interact with phonology (Shankweiler et al., 1995; Rispens et al., 2014). In such processes, the

selection between allomorphs depends on the phonological characteristics of the stem. For example, the selection of the /*t*/, /*d*/ or /*id*/ allomorph in English past tense verb inflection, as in *bake* – *baked*, *try* – *tried* and *bait* – *baited*, depends on the final phoneme of the verb (e.g. Joanisse & Seidenberg, 1998; Joanisse et al., 2000). Thus, problems with the processing of phonological information, such as the verb stem, may affect the acquisition of associated morphological patterns (i.e. verb inflection; Joanisse & Seidenberg, 1999). As for difficulties with syntactic structures, these have been linked to limitations in phonological short-term memory in dyslexia (see Melby-Lervåg et al., 2012; Snowling & Melby-Lervåg, 2016, for meta-analyses): if the processing and storage of phonological information is impaired or limited, this is likely to affect syntactical processing of spoken language. In support of this idea, Robertson and Joanisse (2010) showed that when memory demands are high, children show poorer syntactic processing of spoken sentences, and this effect is more pronounced in children with dyslexia than in TD children.

Alternatively, the grammatical difficulties may be explained through an underlying deficit in statistical learning ability (Nicolson & Fawcett, 2007; 2011; Gabay et al., 2015; Ullman et al., 2019). According to this theory, the grammatical problems observed in dyslexia are not the result of a phonological deficit, but instead the range of impairments associated with dyslexia follow from a domain-general deficit in statistical learning. Generally speaking, sensitivity to regularities and patterns in the (linguistic) input is thought to support the rule-based aspects of language acquisition, including morphology and syntax (e.g. Bannard, Lieven, & Tomasello, 2009; Evans et al., 2009; Kidd & Kirjavainen, 2011; Ullman & Pierpont, 2005). For example, the acquisition of nonadjacent patterns in language, such as the relationship between auxiliaries and inflections on the verb (e.g. the boy *is* runn*ing*, where the intervening verb may vary), may be supported by a mechanism that enables the tracking of their co-occurrence statistics (e.g. Goméz, 2002). This hypothesized relationship between statistical learning and grammatical performance is supported by research that has shown that performance on statistical learning tasks is related to grammatical abilities in TD children. Studies have established such relationships between statistical learning and syntactic priming (Kidd, 2012), grammatical processing (Clark & Lum, 2017) and the comprehension of complex syntactical structures such as passives and relative clauses (Kidd & Arciuli, 2016). Likewise, individual differences in the statistical learning ability of adults has been shown to correlate with the

comprehension of relative clauses (Misyak et al., 2010) and the comprehension of written sentences in general (Misyak & Christiansen, 2012). Moreover, studies have demonstrated impaired statistical learning in children with DLD, who are known to experience grammatical difficulties (see Lammertink et al., 2017, and Lum, Conti-Ramsden, Morgan, & Ullman, 2014, for meta-analyses). No studies to date have explored the relationship between grammatical performance and statistical learning ability in individuals with and without dyslexia.

6.1.3 The current study

In the present study, we tested grammatical abilities of 100 school-aged Dutch-speaking children with and without dyslexia. This was done using two standardized tests of grammar, that target different levels of grammatical knowledge: inflectional morphology and syntax (the "word structure" and "recalling sentences" subtests of the Dutch version of the CELF [CELF-IV-NL]; Kort et al., 2008). Furthermore, we aimed to highlight specific areas of difficulty through an analysis of error patterns. Most importantly, we tested two accounts of dyslexia that make predictions about the relationship between grammar on the one hand and underlying problems in either phonological memory or statistical learning ability on the other hand. Thus, we aimed to answer the following three research questions:

(1) Do children with dyslexia perform worse than their TD peers on grammar as measured with standardized tests (CELF word structure and CELF recalling sentences)?
(2) Do children with dyslexia make different errors than their TD peers on the CELF word structure and/or CELF recalling sentences?
(3) Do phonological memory and/or statistical learning ability contribute to individual differences on the CELF word structure and/or CELF recalling sentences? And, if so,
    a) is this contribution different for the dyslexia versus the control group?
    b) is this contribution different for the CELF word structure versus the CELF recalling sentences?

In relation to research question 3, we focused on measures of phonological memory, since individuals with dyslexia are typically impaired in this area (Melby-Lervåg et al. 2012; Snowling & Melby-Lervåg, 2016) and phonological memory is theorized to contribute to grammatical abilities (e.g. Robertson & Joanisse, 2010). Digit span forward and nonword repetition (NWR) tasks were used to assess phonological short-term memory (i.e. immediate recall), while the digit span backwards was used as a measure of phonological working memory (i.e. the manipulation of phonological information prior to recall; e.g. Baddeley, 2012; Alloway, Gathercole, Kirkwood, & Elliott, 2009). Naturally, these memory tasks also rely on the processing of phonological information and (already established) phonological representations, which is especially true for nonword repetition (Rispens & Baker, 2012).

Statistical learning was tested using two experimental tasks that targeted different aspects of the domain-general ability to detect statistical regularities: visuo-motor sequence learning (serial reaction time [SRT] task) and the learning of auditorily presented nonadjacent dependencies (auditory nonadjacent dependency learning [A-NADL] task). Both statistical learning measures have previously been related to grammatical performance in children and/or have demonstrated impaired learning ability in children with DLD (SRT: e.g. Kidd, 2012; Clark & Lum, 2017; A-NADL: e.g. Iao et al., 2017; Lammertink et al., 2019a). Besides phonological memory and statistical learning measures, our regression analysis includes other potential sources of variance in grammatical performance (children's age, gender, and socio-economic status [SES], and their scores on measures of non-verbal reasoning, vocabulary, and sustained attention).

It is important to note here that any statistical analyses done in order to answer research question 2 are exploratory, since the tasks used to measure grammatical performance were not designed for error analysis specifically. The results from these analyses may further our understanding of the grammatical problems associated with dyslexia and may thereby serve to highlight potentially interesting directions for future research. Moreover, it should be noted that group comparisons regarding statistical learning ability in the present sample have already been discussed in detail elsewhere (van Witteloostuijn et al., 2019, see chapter 4).

6.2 Methods

6.2.1 Participants

The ethical committee of the Faculty of Humanities of the University of Amsterdam approved this study. One-hundred 8- to 11-year-old children were included in the final sample: 50 children with a prior diagnosis of dyslexia (26 girls, 24 boys, mean age in years;months = 9;10) and 50 individually age-matched TD children composed the control group (24 girls, 26 boys, mean age = 9;8). To confirm participation as (non-)dyslexic, word (*Een Minuut Test*; Brus & Voeten, 1972) and pseudo-words (*Klepel*; van den Bos et al., 1994) reading tests were administered. All children with dyslexia in the final sample had a maximum norm score of 6 (i.e. 10th percentile) on word and pseudo-word reading, while TD children had a minimum norm score of 8 (i.e. 25th percentile). Ten additional children with dyslexia and four additional TD children did not meet these pre-determined inclusion criteria regarding their reading scores and were therefore excluded from the final sample. Parental (in the case of dyslexia) and teacher (in the case of control) reports confirmed that all 100 participants in the final sample were native speakers of Dutch and none had diagnoses of (other) developmental disorders such as DLD. All participants were able to complete each of the tasks included in the present study. Please note that, as previously mentioned, the sample described here is identical to the sample described by van Witteloostuijn et al. (2019, see chapter 4; under review, see chapter 5), which focusses on group comparisons on statistical learning measures. Similarly, the control group partly overlaps with studies examining language and statistical learning in children with DLD (Lammertink et al., 2019a; 2019b; 2020). These previous reports thus have a different focus than this study and there is no overlap in interpretation of the data.

Table 6.1. Children with and without dyslexia's mean (and *SD*) age and SES, and results from reading, spelling, nonverbal reasoning, and sustained attention: raw and standardized scores.

	Dyslexia (*N* = 50)		**Control (*N* = 50)**	
	Raw	Standardized	Raw	Standardized
Age	9;10 (0;9)	N/A	9;8 (0;10)	N/A
SES	0.2 (1.2)	N/A	0.2 (1.1)	N/A
Reading words[a]	34.1 (11.7)	3.3 (2.1)	66.3 (11.6)	10.5 (2.2)
Reading pseudo-words[a]	22.0 (8.0)	4.4 (1.6)	61.0 (14.4)	11.1 (2.2)
Spelling[b]	8.4 (4.6)	11.8 (13.7)	18.6 (4.7)	49.9 (24.7)
Nonverbal reasoning[b]	37.2 (6.6)	55.7 (25.0)	37.3 (8.1)	60.1 (28.1)
Vocabulary[b]	117.3 (9.7)	54.8 (21.3)	118.0 (9.4)	58.7 (20.6)
Sustained attention[a]	7.0 (2.5)	7.4 (3.3)	7.8 (1.8)	9.1 (3.0)

Note: Age in years;months. Data regarding SES by postal codes was obtained from the *Netherlands Institute for Social Research* (NISR). Raw scores on reading words (*Een minuut test*, Brus & Voeten, 1972) and pseudo-words (*Klepel*, van den Bos et al., 1994) represent the number of words read within the time limit of 1 and 2 minutes respectively. Raw scores on spelling represent the number of words spelled correctly out of 30 in a Dutch dictation test (Braams & de Vos, 2015). Nonverbal reasoning was measured using *Raven's Standard Progressive Matrices* (Raven & Raven, 2003); raw scores represent the number of items answered correctly out of 60. The *Peabody Picture Vocabulary Test* (PPVT-III-NL; Dunn et al., 2005) was used as a test of receptive vocabulary; raw scores represent the number of items answered correctly out of a maximum of 204 items. Finally, sustained attention was assessed by the *Score!* subtest of the Dutch *Test of Everyday Attention* (*TEA-Ch;* Schittekatte et al., 2007); raw scores represent the number of items answered correctly out of 10. Standardized scores represent either [a] norm scores (norm = 10) or [b] percentile scores (norm = 50).

In line with their diagnosis, children with dyslexia were found to perform significantly worse than the TD children on reading words ($t = 13.83$, $p = 9.1 \cdot 10^{-25}$), reading pseudo-words ($t = 16.75$, $p = 1.6 \cdot 10^{-30}$), and spelling ($t = 11.41$, $p = 1.1 \cdot 10^{-19}$). No evidence for a difference between the two child groups was found regarding their age ($t = .839$, $p = .40$), socio-economic status (SES; $t = .173$, $p = .86$), vocabulary ($t = .367$, $p = .71$), or nonverbal reasoning ($t = .041$, $p = .97$). Children with dyslexia scored lower than TD children on our measure of sustained attention, although this effect did not reach significance ($t = 1.939$, $p = .055$). Table 6.1 above provides more information regarding the measures used

and presents their descriptive results (please note that these results overlap with those presented in van Witteloostuijn et al., 2019, see chapter 4; under review, see chapter 5). Individual differences in age, SES, vocabulary, and nonverbal reasoning are included as control predictors in our regression model that investigates the contribution of phonological memory and statistical learning ability to grammatical performance (research question 3).

6.2.2 Materials

6.2.2.1 Measures of grammatical performance

Children's grammatical abilities were assessed through two subtests of the Dutch version of the standardized CELF (CELF-4-NL; Kort et al., 2008): the CELF word structure and CELF recalling sentences subtests. The CELF word structure task is set up to measure children's ability to apply word formation rules (i.e. inflectional morphology), while the CELF recalling sentences task aims to test children's ability to listen to and repeat sentences, thereby considering grammatical performance at different levels (i.e. morphology, and syntax).

In the CELF word structure task, children were shown pictures and were instructed to finish sentences read out by the experimenter. The task consists of 30 items that are divided into categories targeting different aspects of morphology including pronouns, nouns (i.e. diminutives and plurals), verbs (i.e. subject-verb agreement, tense, and compound verbs), and adjectives (i.e. comparatives and superlatives). Responses were coded as either correct or incorrect, with a maximum score of 30. Children's scores on the CELF word structure task were not converted to standardized (i.e. norm) scores, since norms are available up until the age of 8 and our sample consists of 8- to 11-year-old children. For this reason, one may expect to find scores close to ceiling performance in the current sample, especially in the control group.

The CELF recalling sentences task required children to repeat sentences of increasing length and complexity as dictated by the experimenter. In accordance with the CELF manual, 8-year-old children repeated a maximum of 31 sentences, while children aged 9 years or older were administered a maximum of 23 sentences (the first eight sentences were not administered). Responses received a score of 3 (0 errors), 2 (1 error), 1 (2 or 3 errors), or 0 (4 or more

errors) and testing was discontinued after five consecutive 0 scores. Children's individual score on the CELF recalling sentences task was the total number of points awarded to the administered sentences.

6.2.2.2 Measures of phonological memory

Phonological processing and phonological short-term and working memory were assessed through two tasks: a digit span task (CELF-4-NL; Kort et al., 2008) and a shortened version of a NWR task (NWR-S; le Clercq et al., 2017). Both the forward and backward digit span task were administered, in which children were required to repeat sequences of digits of increasing length either in the same order (forward digit span; 16 items) or in the reversed order (backward digit span; 14 items). In the NWR-S, 22 pre-recorded nonwords were played one at a time and children had to listen carefully and repeat each nonword as accurately as possible. Items in the digit span and NWR-S tasks were scored as either correct or incorrect.

6.2.2.3 Measures of statistical learning

In the SRT task, a visual stimulus continuously appeared in one of four marked locations on a tablet screen. Children were required to press corresponding buttons on a gamepad as accurately and as quickly as they could and started with a practice block (block 1; 28 trials). Without the participants' knowledge, stimuli were presented following a pre-determined sequence (4, 2, 3, 1, 2, 4, 3, 1, 4, 3) in sequence blocks 2–5 and sequence block 7, and was repeated six times per block (i.e. 60 trials). In the intervening block (i.e. disruption block 6), the presentation of the stimulus no longer followed the sequence, but the order of appearance was random for 60 trials. Learning in the SRT task is measured as the increase in reaction times (RTs) in disruption block 6 as compared to the surrounding sequence blocks.

In the A-NADL task, children listened to an artificial language and, unbeknownst to them, their sensitivity to two nonadjacent dependencies was measured. Such nonadjacent dependencies mirror those found in natural languages; for example, the morphosyntactic relationship between auxiliaries and inflections on the verb in English (e.g. "*is* walk*ing*", where the is–*ing* relationship

is nonadjacent and the intervening verb may vary). The task was modelled on the SRT task and thus contained blocks in which the artificial language adhered to these nonadjacent dependency rules (i.e. rule blocks 1–3 and 5) and an intervening block in which it no longer followed these rules (i.e. disruption block 4). Both nonadjacent dependencies *tep X lut* and *sot X mip* had an *a X b* structure, where *a* predicts *b* and the intervening *X* was selected from a set of 24 two-syllable nonwords, and both dependencies were presented 24 times per rule block. Besides the items containing the nonadjacent dependency rules, each rule block contained 12 filler trials with f_1 *X* f_2 structure where f_1 does not predict f_2 (f_1 and f_2 are taken from a set of 24 one-syllable nonwords that do not include *tep*, *sot*, *lut* or *mip*, and *X* refers to the same set of 24 two-syllable nonwords used in the *a X b* structure). In the intervening disruption block, the occurrence of *lut* and *mip* was no longer predicted by the *a X b* rule: in 24 out of 30 trials, *lut* and *mip* still occurred in the *b* position, but one of the one-syllable fillers *f* occurred in the *a* position (i.e. *f X b* structure). The remaining 6 trials were entire filler items (i.e. f_1 *X* f_2 structure). Children performed a word-monitoring task in which they tracked the occurrence of one of the two predictable nonwords (i.e. the *b* element in the *a X b* structure). Half of the participants was assigned to *lut* as a target and half to *mip*. Children were instructed to press a green button when they heard the target nonword and to press a red button when they did not hear the target nonword (Lammertink, van Witteloostuijn et al., 2019; López-Barroso et al., 2016). As in the SRT task, learning in the A-NADL task is reflected by slower RTs to input in the disruption block than to rule-governed input in the surrounding rule blocks.

In both statistical learning tasks, accuracy and RTs to each trial were recorded. As explained, learning is evidenced by shorter RTs to structured input as compared to random input and, therefore, the individual measures of learning used in the regression analysis are difference scores (SRT: normalized RT in disruption block 6 minus mean normalized RT in sequence blocks 5 and 7, A-NADL: normalized RT in disruption block 4 minus mean normalized RT in rule blocks 3 and 5).

6.2.3 General procedure

As mentioned in §6.2.1, children in the present study were tested as part of a larger project investigating statistical learning and its relation to language in children with and without dyslexia and DLD (van Witteloostuijn et al., 2019, see chapter 4; Lammertink et al., 2019a; 2019b; 2020). A test battery including the tasks reported on here was administered one-on-one by an experimenter in the child's home or school. The test battery took approximately three hours to complete and was divided into three testing sessions. Importantly, each testing session consisted of one statistical learning measure, combined with a range of background and language measures. The orders between and within sessions were counter-balanced; and children were randomly assigned to one out of six testing orders.

The CELF word structure, CELF recalling sentences, PPVT, and digit span tasks were dictated by the experimenter. Instructions (SRT, A-NADL) and auditory stimuli (A-NADL, NWR-S) in the statistical learning and NWR-S tasks were pre-recorded by a native Dutch speaker and were played over Sennheiser HD 201 headphones. PPVT images were shown on a Windows Surface 3 tablet. The SRT and A-NADL tasks were programmed and administered through E-prime 2.0 and displayed on the same tablet (Psychology Software Tools, 2012; Schneider et al., 2012). Accuracy and RTs in the SRT task were logged using a Trust wired GXT540 gamepad controller, while responses to the A-NADL task were logged through an external button box. Verbal responses in the CELF word structure, CELF recalling sentences and NWR-S tasks were recorded using an Olympus DP-211 voice recorder.

6.2.4 Scoring and analysis

The following sections provide details of our method of scoring and analyses regarding group comparisons, the error exploration and the regression model. All analyses were performed in *R* software (*R* core team, 2019); raw (summary-level) data files and *R* Markdown and html files containing all analyses reported in the present study can be found on our Open Science Framework (OSF) project page (https://osf.io/kjctf/).

6.2.4.1 Group comparisons

Individual *t*-tests were run on children's raw scores on our outcome measures (CELF word structure and CELF recalling sentences; research question 1) and on the raw scores on the tests assessing phonological memory (digit span and NWR-S) in order to investigate whether a difference in performance is observed between participants with and without dyslexia. As mentioned, investigations of group differences on the statistical learning measures (SRT and A-NADL) were already reported on elsewhere (van Witteloostuijn et al., 2019, see chapter 4), and are thus not re-analysed here, although these findings are summarized in §6.3.1.

6.2.4.2 Exploratory error explorations

Children's performance on the CELF word structure and CELF recalling sentences was examined in more detail through an error analysis. This allowed us to explore whether children with dyslexia make qualitatively different errors than their TD peers (research question 2). Since items of the CELF word structure are already divided into categories, we inspected the total number of errors (and proportion of answers correct) per category (see §6.2.2.1). To explore potential differences between children with and without dyslexia in their performance on the CELF word structure categories, *t*-tests were run on the number of errors.

For ease of error analysis, responses to the CELF recalling sentences were recoded as either correct or incorrect (instead of 0, 1, 2, or 3; see §6.2.2.1). We categorized errors according to a pre-determined scoring schedule. Different types of errors were distinguished: errors pertaining to the inflectional morphology of verbs (subject-verb agreement, tense, overgeneralization, and lexical errors) and nouns (plural, article choice, and lexical errors), and errors with the referential use of pronouns (demonstratives). These error categories combined will be referred to as "specific errors". The remaining errors (i.e. errors that could not be categorized under specific error categories) were deemed "unspecific errors", which included omissions, additions, replacements and displacements of words that we did not analyse further (e.g. uttering a word in a different position in the sentence or switching two words). As in the CELF word structure analysis, *t*-tests were run to investigate potential group differences on

the number of specific errors in certain categories. Furthermore, to inspect the level of syntax, 19 sentences were marked as "syntactically complex"; these consisted of passive sentences (N = 6) and sentences containing a subordinate clause (N = 13). Moreover, the length of sentences was also considered in our analyses of the CELF recalling sentences. The effects of syntactic complexity, sentence length and group (dyslexia versus control) were explored using a generalized linear mixed effects (GLMER) model built using the *lme4* package for *R* (version 1.1-13; Bates et al., 2014). 95% CIs were computed through Wald's approximation for confidence intervals (CIs) and raw sentence length (i.e. number of words in target sentence) was centered and scaled by standard deviation. The categorical predictors included in the model were sentence complexity and group, which were orthogonally contrast-coded. Sentence complexity was coded into two contrasts such that the first contrasted simple (coded as - $^2/_3$) and complex sentences (passive and subordinate, coded as + $^1/_3$) and the second contrasted the two complex sentence types (i.e. passive coded as - $^1/_2$ and subordinate coded as + $^1/_2$). The coding of Group was identical to the coding reported for the regression model: the control group was coded as - $^1/_2$ and the dyslexia group as + $^1/_2$. The random effect structure of the model contained by-subject intercepts and by-subject random slopes for sentence length, sentence complexity and the interaction between sentence length and sentence complexity.

Since testing on the CELF recalling sentences was discontinued after five consecutive 0 scores (see §6.2.2.1), it is important to note that testing was halted after a similar number of sentences in both participant groups of children with and without dyslexia (dyslexia: 29.5 [SD = 2.8], TD: after 30.2 [SD = 2.0]). Although a subset of children (i.e. children over the age of 8) did not complete sentences 1 through 8, we disregard this in our error analyses since children were individually matched on age. Of all 2320 sentences administered to our 100 participants, 10 sentences resulted in null responses that were categorized as missing data and were excluded from analyses (dyslexia: N = 6, control: N = 4).

6.2.4.3 Regression analysis

We set up a linear regression model to examine whether a range of predictors contribute to individual differences in performance on our outcome measures of

grammar (research question 3). This was done using the "lm" function included in *R*, which modelled grammatical performance by a number of control predictors (age, gender, SES, nonverbal reasoning, vocabulary, and attention) and predictors relevant to research question 3 (phonological memory: digit span and NWR-S, statistical learning: SRT, and A-NADL). Group membership (dyslexia versus TD) was added as a predictor in order to assess interactions between group and other predictors (research question 3a). The significance of predictors to both grammatical measures combined (CELF word structure and recalling sentences) was determined through the "Manova" function in the *car* package for *R* (version 2.1-5; Fox et al., 2012). The effects of phonological memory and of statistical learning on grammar performance were investigated by comparing the full model to models from which both measures assessing phonological memory (digit span and NWR-S) and both statistical learning measures (SRT and A-NADL) were removed. To compare the contribution of predictors to CELF word structure versus CELF recalling sentences (research question 3b), we computed 95% CIs using the profile method ("confint" function in *R*) and examined the overlap of 95% CIs of individual predictors for the two measures of grammar performance. Importantly, raw scores on continuous outcome variables (CELF word structure and CELF recalling sentences) and predictors (age, SES, non-verbal reasoning, attention, PPVT, digit span, NWR-S, SRT and A-NADL) were centered and scaled. The categorical predictors, i.e. Gender and Group, were orthogonally contrast-coded: females were coded as - $^1/_2$ and males as + $^1/_2$, and, similarly, the control group was coded as - $^1/_2$ and the dyslexia group was coded as + $^1/_2$.

6.3 Results

We present the results regarding our three research questions: the group comparisons pertaining to research question 1 are described in §6.3.1, followed by the error exploration in §6.3.2 (research question 2), and the regression analysis in §6.3.3 (research question 3). While the analyses related to research questions 1 and 3 can be viewed as confirmatory, the analyses related to research question 2 are exploratory in nature. Additionally, our regression analysis provides us with some exploratory findings that may be of interest and are reported separately following recommendations by Wagenmakers et al. (2012).

As mentioned, files containing all data and analyses presented here can be found on our OSF project page (https://osf.io/kjctf/).

6.3.1 Group comparisons

Table 6.2 presents the mean (and *SD*) scores on the two measures of grammar (CELF word structure and CELF recalling sentences), and the phonological memory (digit span, NWR-S) and statistical learning (SRT, A-NADL) measures included as predictors in our regression model that is discussed in §6.3.3 (see also van Witteloostuijn et al., 2019, see chapter 4; under review, see chapter 5). In order to answer our first research question, we examined group effects on children's grammatical performance as measured by the CELF word structure and recalling sentences subtests. Results reveal that participants with dyslexia achieved significantly lower scores on the CELF word structure (t = 3.082, p = .0027). The children with dyslexia also achieved lower scores on the CELF recalling sentences, but this effect did not reach significance (t = 1.957, p = .053). Out of 50 children with dyslexia, 9 received a norm score of 6 (i.e. 10th percentile) or lower on the CELF recalling sentences, while 7 out of 50 TD children received a norm score of 6 or lower. No norm scores are available for the CELF word structure subtest (see §6.2.2.1).

Furthermore, the children with dyslexia performed significantly worse than the TD children on the digit span forward task (t = 5.36, p = 5.5·10^{-7}) and the NWR-S (t = 3.962, p = .00014), which both assess phonological processing and short-term memory. No evidence of such a difference between participant groups was found for the digit span backward that targets phonological processing and working memory (t = 1.257, p = .21). Finally, as previously published in van Witteloostuijn et al. (2019, see chapter 4), evidence of learning was found for both statistical learning measures when looking at children with and without dyslexia together. Importantly, however, no evidence of a group difference emerged for either the SRT (p = .61) or the A-NADL (p = .87) task.

Table 6.2. Mean (and *SD*) scores on measures of grammar, phonological skills, and statistical learning: raw and standardized scores per group.

	Dyslexia (*N* = 50)		**Control (*N* = 50)**	
	Raw	Standardized	Raw	Standardized
CELF word structure[c]	26.9 (1.8)	N/A	27.9 (1.3)	N/A
CELF recalling sentences[a]	55.2 (10.3)	8.5 (2.2)	59.7 (12.9)	9.9 (2.9)
Digit span forward[a]	7.3 (1.5)	7.7 (2.6)	8.9 (1.5)	10.7 (2.9)
Digit span backward[a]	4.2 (1.1)	9.0 (2.5)	4.5 (1.5)	10.0 (3.2)
NWR-S[c]	7.3 (2.7)	N/A	9.7 (3.3)	N/A
SRT[c]	0.29 (0.28)	N/A	0.27 (0.27)	N/A
A-NADL[c]	0.15 (0.33)	N/A	0.17 (0.28)	N/A

Note: Raw scores: CELF word structure = number of items correct out of 30, CELF recalling sentences = total score on administered sentences, Digit span = number of items answered correctly out of 16 (forward) or 14 (backward), NWR-S = number of nonwords repeated correctly out of 22, SRT = difference in normalized RTs (RT disruption – RT sequence), A-NADL = difference in normalized RTs (RT disruption – RT rule). Standardized scores represent either [a] norm scores (norm = 10) or [b] percentile scores (norm = 50). [c] No standardized scores are available for the CELF word structure, NWR-S, SRT and A-NADL tasks.

6.3.2 Error exploration

6.3.2.1 Word structure

Overall, the children with dyslexia made an average of 3.1 errors out of 30 (range: 0–10 errors) and the TD children made 2.1 errors (range: 0–5 errors; see Figure 6.1). Performance on two categories of the CELF word structure task was found to be at ceiling both in participants with and without dyslexia: regular plurals (dyslexia: 3 errors, 98.5% accuracy, TD: 1 error, 99.5% accuracy) and past tense formation (dyslexia: 3 errors, 98.5% accuracy, TD: 2 errors, 99% accuracy). Moreover, on categories eliciting demonstrative and personal pronouns, the children with and without dyslexia achieved comparable levels of accuracy (demonstrative pronouns: dyslexia 32 errors, 68% accuracy, TD 34 errors, 66% accuracy, personal pronouns: dyslexia 26 errors, 87% accuracy, TD: 23 errors, 88.5% accuracy). The other categories may inform us about different error patterns in children with dyslexia as compared to their TD peers, as participants

with dyslexia achieve numerically lower accuracy levels. These categories include irregular plurals (dyslexia: 17 errors, 91.6% accuracy, TD: 7 errors, 96.5% accuracy), diminutives (dyslexia: 18 errors, 92.8% accuracy, TD: 9 errors, 96.4% accuracy), compound verbs (dyslexia: 46 errors, 54% accuracy, TD: 28 errors, 72% accuracy), and comparative superlatives (dyslexia: 10 errors, 96% accuracy, TD: 2 error, 99.2% accuracy). The difference in accuracy levels between participants with and without dyslexia reaches significance on irregular plurals and compound verbs (t = -2.148, p = .034 and t = -3.368, p = .0011 respectively). Differences between the children with and without dyslexia do not reach significance on diminutives or comparative superlatives (t = -1.718, p = .089 and t = -1.871, p = .065 respectively).

Closer inspection of the error pattern on the irregular plurals (4 items) reveals that the children with dyslexia make most errors on the item *ei – eieren* [ɛi – ɛiərə(n)] ('egg – eggs'; 12 errors), followed by *schip – schepen* [sxɪp – sxeːpə(n)] ('ship – ships'; 5 errors), and fewest errors are made on *koe – koeien* [ku – kujə(n)] ('cow – cows'; 2 errors) and *glas – glazen* [ɣlɑs – ɣlaːzə(n)] ('glass – glasses'; 2 errors). The errors in the TD participants are distributed more equally (*koe – koeien*: 1 error, *ei – eieren*: 2 errors, *schip – schepen*: 3 errors; *glas – glazen*: 0 errors). Generally speaking, errors on irregular plurals are cases of overgeneralization: participants apply the regular plural rules (add /ə(n)/ or /s/) to irregular nouns (i.e. *ei – *eien* *[ɛiə(n)], *schip – *schippen* *[sxɪpə(n)], *glas – *glassen* *[ɣlɑsə(n)], *koe – *koes* *[kus], instead of applying the required more complex suffix (/ərə(n)/ as in *ei – eieren*) or alteration of the noun stem (i.e. vowel lengthening as in *schip – schepen* [sxɪp – sxeːp-ə(n)] and *glas – glazen* [ɣlɑs – ɣlaːz-ə(n)], or stem alteration as in *koe – koeien* [ku – kuj-ə(n)].

Regarding compound verbs, the majority of errors (dyslexia: 40 out of 46; TD: 27 out of 28) is made on the item *wassen af* ('[they are] washing the dishes') and only few errors are made on the item *speelt gitaar* ('[he/she] plays the guitar'). These errors are cases in which the child fails to separate the two verbal elements, such as *zij zijn aan het afwassen* ('they are washing the dishes'; this is not ungrammatical, but may be an avoidance strategy), and/or cases in which the infinitive form of the verb is used (i.e. **zij gitaar spelen* ['she guitar plays'], or **zij afwassen* ['they washing the dishes']).

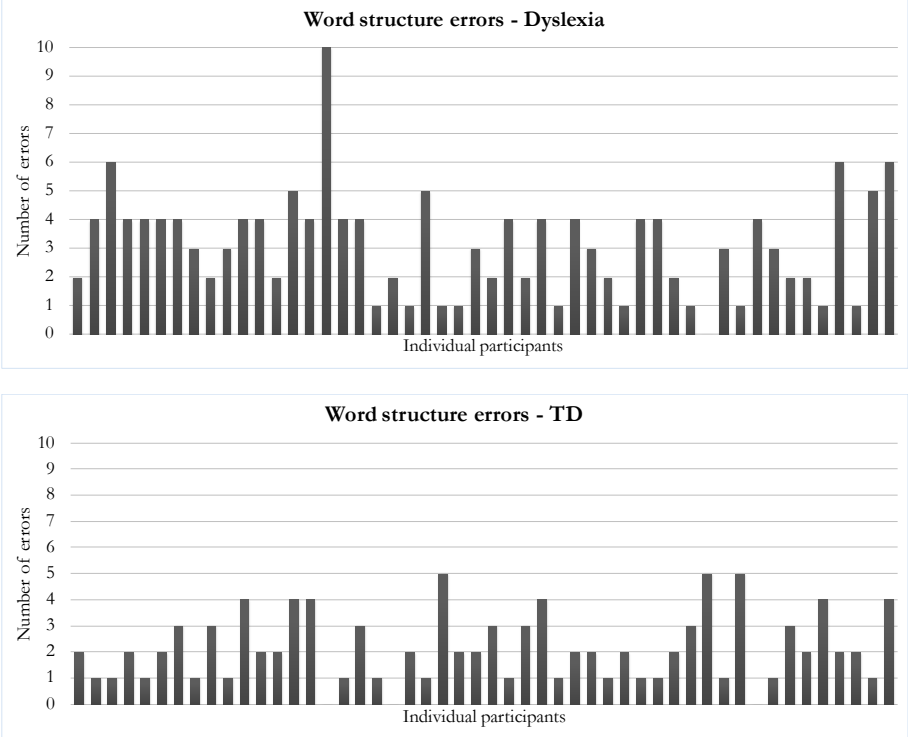


Figure 6.1. Histogram showing the distribution of performance on the CELF word structure subtest; children with dyslexia are presented in the top graph, TD children are presented in the bottom graph. Each bar represents the number of errors (out of 30 test items) of an individual participant.

6.3.2.2 Recalling sentences

In our GLMER model predicting children's performance on the CELF recalling sentences task, we found a significant effect of sentence length: accuracy was lower for longer sentences than for shorter sentences (odds ratio estimate = a factor of 1.5 per standard deviation, 95% CI = [1.4 … 1.6], $p = 3.5 \cdot 10^{-35}$). There was also a significant effect of sentence complexity; accuracy was lower for sentences that contained a complex syntactical structure as compared to simple sentences (odds ratio estimate = 58, 95% CI = [13 … 261], $p = 1.1 \cdot 10^{-7}$). As may be expected, these two predictors were found to significantly interact with one another, indicating that the effect of length on performance is stronger in complex sentences than in simple sentences (odds ratio estimate = 1.4, 95% CI

= [1.2 … 1.5], $p = 7.5 \cdot 10^{-6}$). Furthermore, accuracy was significantly lower on sentences containing subordinate clauses as compared to passive sentences (odds ratio estimate = 23, 95% CI = [5 … 119], $p = .00015$) overall and the effect of length was found to be significantly stronger for subordinate clauses than for passive sentences (odds ratio estimate = 1.5, 95% CI = [1.3 … 1.8], $p = 7.9 \cdot 10^{-8}$). The effect of group in interaction with sentence complexity (odds ratio estimate = 1.4, 95% CI = [0.07 … 29], $p = .81$) or sentence length (odds ratio estimate = 1.0, 95% CI = [0.9 … 1.1], $p = .87$) is non-significant. Similarly, none of the three-way interactions with group are found to be significant. Thus, performance on the CELF recalling sentences task is influenced by sentence length and sentence complexity in children, and we find no evidence of a difference in performance between children with and without dyslexia regarding the effects of sentence length and sentence complexity.

A total number of 4762 errors (dyslexia: $N = 2525$, TD: $N = 2237$) on the CELF recalling sentences task were classified into pre-determined categories (see §6.2.4.2). Further inspection of these errors reveals that the largest proportion of total errors are classified as non-specific: omissions of words (dyslexia: 35.5%, TD: 36.4%), replacements, displacements, and switches of words (dyslexia: 26.9%, TD: 25.6%), and additions of words (dyslexia: 11.1%, TD: 11.9%). Together, these error categories account for approximately 74% of all errors made on the CELF recalling sentences task. Both for the children with and without dyslexia, non-specific errors involve function words slightly more often than content words (dyslexia: 51.2% and 48.8% of non-specific errors respectively, TD: 51.9% and 48.1% of non-specific errors respectively).

As explained in §6.2.4.2, the remaining 26% of errors (dyslexia: $N = 671$, TD: $N = 585$) were labelled as specific errors and were divided into errors pertaining to nouns (plurals, article choice, and lexical selection errors), verbs (subject-verb agreement, tense, overgeneralization, and lexical errors), and demonstrative pronouns. Table 6.3 presents a summary of these results. Of these specific errors, the children with dyslexia made an average of 13.4 errors (range: 1–24 errors) and the TD children made 11.7 errors (range: 1–26 errors; see Figure 6.2).

Table 6.3. Recalling sentences: numbers of errors in the categories covering verb, noun and pronoun errors per group.

		Dyslexia (*N* = 50) # Errors	**TD (*N* = 50)** # Errors	
Verbs	Subject-verb agreement	15	23	38
	Tense	121	107	228
	Overgeneralization	13	4	17
	Lexical	192	192	384
Nouns	Plural	6	7	13
	Article choice	36	30	66
	Article choice (definite)	67	41	108
	Lexical	168	139	307
Pronouns	Demonstrative	53	42	95
Total		671	585	1256

The largest proportion of specific errors was classified as lexical errors, both in the children with dyslexia (verbs: N = 192, nouns: N = 168) and in the TD children (verbs: N = 192, nouns: N = 139), corresponding to approximately 55% of specific errors. Regarding nouns, the children make very few pluralization errors (dyslexia: N = 6, TD: N = 7). More errors are made concerning article choice: both the choice between indefinite and definite articles (*een* versus *de*/*het*; dyslexia: N = 36, TD: N = 30) and between the two definite articles (*de* versus *het*; dyslexia: N = 67, TD: N = 41). Exploratory *t*-tests suggest that children with dyslexia may make more errors regarding the choice between the two definite articles in Dutch (t = 2.039, p = .044). No evidence of a difference between groups is found for errors concerning the choice between indefinite and definite articles (t = 1.178, p = .24). Secondly, regarding verbal morphology, the children made a small number of subject-verb agreement errors (dyslexia: N = 15, TD: N = 23) and overgeneralization errors (dyslexia: N = 13, TD: N = 4), whereas tense errors are more frequent (dyslexia: N = 121, TD: N = 107). No evidence of a difference in performance between children with and without dyslexia is found regarding the number of subject-verb agreement errors (t = -0.906, p = .37) and tense errors (t = 1.012, p = .31). The children with dyslexia were found to produce significantly more verb overgeneralization errors (t = 2.411, p = .019). These are instances where children apply the regular Dutch past tense rule (i.e. add /*te*/) to irregular verbs, such as *koop* – **koopte* (correct: *kocht*; 'buy – *buyed – bought').

However, please note the low number of errors in this category overall. Lastly, no evidence of a difference in performance between children with and without dyslexia is found regarding the incorrect use of the demonstrative pronoun (dyslexia: $N = 53$, TD: $N = 42$; $t = 1.559$, $p = .12$).

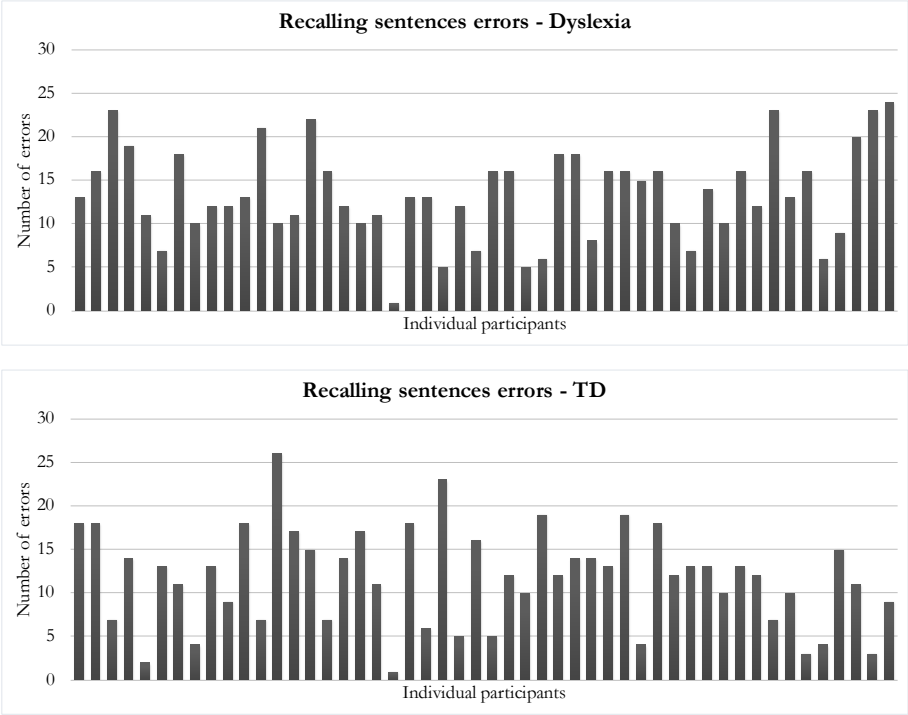


Figure 6.2. Histogram showing the distribution of performance on the CELF recalling sentences subtest; children with dyslexia are presented in the top graph, TD children are presented in the bottom graph. Each bar represents the number of specific errors of an individual participant.

6.3.3 Regression analysis

6.3.3.1 Regression analysis: confirmatory findings

In order to answer research question 3, we performed a linear regression analysis to investigate the effects of phonological memory (digit span and NWR-S) and statistical learning (SRT and A-NADL) on children's performance on the CELF

word structure and recalling sentences subtests. Manova results show that NWR-S (Wilk's λ= .88, F[2,80] = 5.28, p = .0070) scores significantly affect grammar performance (CELF word structure and recalling sentences combined). The effects of the digit span forward (Wilk's λ= .96, F[2,80] = 1.89, p = .16) and backward (Wilk's λ= .99, F[2,80] = 0.22, p = .81) tasks do not reach significance. Importantly, when we compare the full model to a model where the phonological memory measures (digit span and NWR-S) are removed, this results in a significant decrease in the fit of the model (F[12,162] = 2.80, p = .0017). The model provides no evidence of an effect of statistical learning on individual differences in performance on the CELF WS and RS (SRT: Wilk's λ= .99, F[2,80] = 0.36, p = .70; A-NADL: Wilk's λ= .99, F[2,80] = 0.29, p = .75). Comparing the full model to a reduced model where the statistical learning measures (SRT and A-NADL) are removed does not reveal a significant difference in fit between the models (F[8,162] = 1.04, p = .41). Thus, we can conclude that phonological memory skills contribute to the grammatical performance of children with and without dyslexia; and we find no evidence for or against the hypothesis that statistical learning contributes to children's grammatical performance.

Regarding potential differences between children with and without dyslexia (research question 3a), the model shows a significant interaction between phonological processing and short-term memory, as measured by the NWR-S, and group (NWR-S*Group: Wilk's λ= .92, F[2,80] = 3.51, p = .035). To follow up on this interaction, we calculated correlations between NWR-S and CELF performance separately for both groups. Results show that the correlation between nonword repetition and grammatical performance is significant in both groups (TD: r = .580, t(48) = 4.934, p = $1.0 \cdot 10^{-5}$; dyslexia: r = .504, t(48) = 4.042, p = .00019). No other interactions with group are found to be significant. As for potential differences in the contribution to CELF word structure versus recalling sentences performance (research question 3b), we cannot conclude that there is a difference in effect of the NWR-S due to overlapping 95% CIs (effect NWR-S on CELF word structure: β = +.20, 95% CI [-.050 ... +.46], p = +.11; effect NWR-S on CELF recalling sentences: β = .28, 95% CI [+.089 ... +.47], p = .0044).

6.3.3.2 Regression analysis: exploratory findings

Besides enabling us to answer research question 3, the regression model provides us with exploratory findings that may be of interest. Firstly, non-verbal reasoning (Raven; Wilk's $\lambda = .87$, $F[2,80] = 6.06$, $p = .0035$) and vocabulary (PPVT; Wilk's $\lambda = .81$, $F[2,80] = 9.28$, $p = .00024$) are found to contribute to CELF word structure and recalling sentences performance combined. Secondly, there is no evidence that the effect of group membership contributes to children's grammatical performance over and beyond the other predictors included in the model (Wilk's $\lambda = .95$, $F[2,80] = 2.31$, $p = .11$).

In addition to these findings provided by the model, we here wish to further explore the effects of phonological memory and statistical learning on grammar performance, since these constructs were measured using multiple tests. Firstly, the results described above suggest that the effect of phonological memory is largely carried by the NWR-S. This is corroborated by a further analysis: removal of digit span forward and backward does not significantly affect the fit of the model ($F[12,162] = 0.56$, $p = .81$). In other words, the digit span tasks do not significantly add to the model above and beyond the NWR-S. Similarly, we wish to explore the effect of the A-NADL task on its own, since this statistical learning task is considered to model aspects of grammar acquisition. Removing the SRT task from the model does not result in a significant decrease in fit ($F[4,162] = 0.90$, $p = .46$), and the effect of the A-NADL task on its own remains non-significant (Wilk's $\lambda = .99$, $F[2,80] = 0.31$, $p = .73$). Please note that the effect of the SRT and A-NADL tasks combined also did not reach significance (§6.3.3.1). Therefore, we find no evidence for (or against) the hypothesis that the A-NADL and the SRT contributes to grammar performance in children with and without dyslexia.

6.4 Discussion

The goal of this study was to examine the performance of Dutch-speaking school-aged children with and without dyslexia on standardized measures of inflectional morphology and syntax. We investigated whether phonological memory and statistical learning ability contributed to children's grammatical

performance, in order to shed light on the underlying causes of the linguistic difficulties associated with dyslexia. Here, we first discuss the findings concerning group and error pattern analyses of tasks assessing inflectional morphology and syntax (research questions 1 and 2), followed by a discussion of the contributions of phonological memory and statistical learning to children's grammatical performance (research question 3).

6.4.1 Grammatical difficulties in children with dyslexia

In line with previous studies examining the performance of children with dyslexia on (standardized) tests of grammar, children with dyslexia in the present study achieved lower scores on the CELF word structure subtest that assesses inflectional morphology (p = .0027; see also Joanisse et al., 2000) and the CELF recalling sentences (p = .053), targeting both morphology and syntax (see also Carroll & Myers, 2010). When investigating the effects of a range of predictors on grammatical performance, results showed that group membership (i.e. having a diagnosis of dyslexia or not) did not contribute to individual differences in grammar scores over and beyond other contributors to performance (e.g. vocabulary, non-verbal reasoning, nonword repetition and digit span). Together, these findings agree with earlier findings that showed that difficulties in the area of grammar in individuals with dyslexia exist, but are likely to be subtle (Rispens et al., 2014), at least in 8 to 11-year-olds.

To explore the nature of the observed difficulties within the CELF word structure and recalling sentences subtests, we performed a fine-grained analysis of children's error patterns. Here, we highlight the most important findings. Firstly, regarding the CELF word structure subtest, no evidence of a difference between participant groups was found on the production of diminutives (i.e. producing the correct diminutive suffix on nouns as in *boom–pje*; 'tree–DIM' [diminutive marker]; see also Boersma, 2018), comparative superlatives (e.g. *snel*, *snel–ler*, *snel–st*; 'fast, fast–er, fast–est'), regular and irregular past tense, or pronouns. Recall that accuracy on demonstrative pronouns was low in both participant groups (dyslexia 68%, control: 66%): children overuse the common demonstrative pronoun *die* ('that') in situations where the neuter pronoun *dat* ('that') is required. The overgeneralization of the common gender in Dutch is a pattern previously described for TD children (see e.g. Blom, Polišenská, &

Weerman, 2008). Interestingly, participants with dyslexia were found to achieve scores close to ceiling performance on items targeting regular plurals (98.5% accuracy), while accuracy was found to be lower than their TD peers on items assessing irregular plurals. Errors on irregular plurals were cases of overgeneralization of the regular plural rule. As suggested by Ullman (2001) in his declarative/procedural model of language, the use of irregulars is thought to be supported by the mental lexicon, while the use of regulars depends on the application of structural rules (i.e. grammar). Thus, in the case of irregular plurals, instead of retrieving the correct (irregular) plural form from their lexical memory (e.g. *ei*, *ei–eren*; 'egg–PL' [plural marker]), participants with dyslexia were more likely to incorrectly apply the regular pluralization rule than TD participants (e.g. *ei*, *ei–*en*; 'egg–*PL'). This pattern of findings may suggest a problem with lexical retrieval in dyslexia, which is in line with previous studies indicating poor performance on tasks assessing lexical retrieval (i.e. rapid automatized naming; e.g. Bexkens, van den Wildenberg, & Tijms, 2014). Furthermore, participants with dyslexia were outperformed by control participants on separable compound verbs. This is an indication of difficulties with production of the correct verb-second word order in Dutch: the finite verb (i.e. the verb that expresses tense and/or agreement) appears in second position (*zij wassen af*, 'they wash up') and, thus, the production of an infinite verb in second position is ungrammatical (**zij afwassen*, *'they washing up'). Problems related to the verb-second phenomenon have previously been observed in children with DLD and are argued to be the result of underlying processing and working memory deficits (e.g. Blom, Vasić, & de Jong, 2014; de Jong, 1999; Rice & Wexler, 1996; Verhoeven, Steenge, & van Balkom, 2011). Moreover, both overregularization and verb-second avoidance strategies are known to occur in the language production of younger TD children and have been proposed to be the result of weak memory traces (e.g. Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992; Wexler, 1994). Thus, likewise, the difficulties with these phenomena in older children with dyslexia may be partially explained by limitations in the retrieval of lexical information.

Secondly, children's sentence recall accuracy was lower when sentences were longer and/or syntactically more complex. More importantly, there was no evidence that these effects of sentence length and syntactic complexity affected children with and without dyslexia differently. Thus, although CELF recalling sentences performance is influenced by both short-term memory load and

syntactic complexity, we find no evidence that this effect is more pronounced in children with dyslexia as previously reported by Robertson and Joanisse (2010) for sentence comprehension. As for specific error types, the children with dyslexia made more definite article selection errors (common *de* or neuter *het*) and produced more past tense overgeneralization errors (applying the regular morphological rule to irregular verbs, e.g. *koop–*te*; 'buy–*ed'). As for the irregular plural errors in the CELF word structure subtest, these error types appear to be lexical in nature. In Dutch, the correct choice between the common and neuter article depends on the lexical knowledge of the noun: since Dutch noun gender is largely arbitrary, it has to be stored in the mental lexicon for each noun separately (e.g. Blom et al., 2008; Orgassa & Weerman, 2008). Thus, we find errors suggesting difficulties in lexical retrieval, both in the CELF word structure and the CELF recalling sentences subtests. Alternatively, the application of regular morphological rules to irregular nouns or verbs might result from so-called "hypercorrection" (de Bree, van der Ven, & van der Maas, 2017). Since dyslexia treatments focus largely on teaching regular morphological rules, children with dyslexia may show a tendency to apply these rules, even in case of exceptions. However, this explanation cannot account for the lexical errors regarding the choice of the correct definite article.

There are a number of limitations that we would like to point out here. Firstly, the exploratory findings presented here should be interpreted with caution. Future research needs to further investigate these findings regarding differences in error patterns between children with and without dyslexia, to test whether the findings reported here are reliable and generalizable. Secondly, the CELF word structure subtest may not have been maximally sensitive to differences in performance in the current sample due to the fact that it is designed to test children between 5 and 8 years of age. Finally, it is worth noting that the results presented here are based on few items (e.g. compound verbs in the CELF word structure) and/or a low number of errors overall (e.g. overgeneralization of the regular past tense in the CELF recalling sentences). Future studies comparing children with dyslexia not only to a group of age-matched TD children, but also to a group of children with DLD and/or with TD groups matched on reading ability, may further our understanding of the extent of grammatical difficulties in dyslexia and of the overlap with the problems observed in DLD.

6.4.2 Contributions to grammar performance in children with and without dyslexia

The primary aim of the present study was to establish whether phonological memory and statistical learning ability contribute to children's performance on the CELF word structure and recalling sentences subtests (i.e. grammatical ability), while controlling for children's age and SES, and scores on tasks measuring their non-verbal reasoning, vocabulary, and attention. We conclude that phonological processing and phonological short-term and working memory contribute to the grammatical performance of children with and without dyslexia, above and beyond other predictors in the model. Thus, the results from our regression analysis are congruent with the idea that grammatical problems observed in dyslexia may be partially explained by an underlying weakness in the area of phonology (e.g. Shankweiler et al, 1995; Joanisse et al., 2000). More specifically, problems with the processing and short-term storage of phonological information, as measured by nonword repetition and digit span tasks, contribute to difficulties in the areas of inflectional morphology and syntax (see also Robertson & Joanisse, 2010). The correct processing and memorization of verbal material is relevant in both the CELF word structure and recalling sentences subtests, since they involve the processing of spoken sentences and either completing (CELF word structure) or repeating (CELF recalling sentences) these sentences. The link between phonological memory and grammar performance in the present study is further supported by the finding that children with dyslexia make more errors than TD children on compound verbs, which has previously been related to a phonological processing and memory limitation in children with DLD (e.g. Blom et al., 2014). Similarly, it is in line with the observation that participants were affected by sentence length in their performance on the CELF recalling sentences. Taken together, these results underline the important role that phonological processing and phonological memory play in grammatical performance, and they suggest that the grammatical problems observed in dyslexia may stem from an underlying problem in the area of phonological processing (e.g. Shankweiler et al., 1995).

We could not conclude whether or not statistical learning ability, as assessed through SRT and A-NADL tasks, contributes to children's grammatical performance. Although statistical learning has been shown to be impaired in

individuals with dyslexia (e.g. Gabay et al., 2015; Lum et al., 2013) and DLD (e.g. Lammertink et al., 2017; Lum et al., 2014) and has been related to grammatical abilities in TD children (e.g. Clark & Lum, 2017; Kidd, 2012; Kidd & Arciuli, 2016), our data do not provide evidence for (or against) the relationship between statistical learning on the one hand and inflectional morphology and syntax on the other hand. While this may seem surprising, other studies of the relationship between statistical learning and language performance have similarly reported null results (e.g. West et al., 2017). Recently, the reliability of statistical learning measures has been questioned, especially in child participants (e.g. Arnon, 2019a; West et al., 2017). Measures that are currently used may not be suitable to examine the hypothesized relationship with linguistic performance (e.g. Arnon, 2019b). However, note that the measures used in the present study were reliable at detecting learning in child participants with and without dyslexia overall. Moreover, as presented in van Witteloostuijn et al. (under review, see chapter 5), the split-half reliability of the SRT task in the present sample was $r = .71$, 95% CI = [.58 … .81]. The split-half reliability of the online measure of the A-NADL task used in the present study, although not calculated for the present sample, was previously reported to be $r = .79$, 95% CI = [.66 … .87], in a sample of 72 7- to 10-year-old children with and without DLD (Lammertink et al., 2019a). Thus, the split-half reliability coefficients of the statistical learning measures used in the present study are relatively high and approach the recommended value of $r = .80$ (see e.g. Nunnally & Bernstein, 1994; Siegelman et al., 2017a; Streiner, 2003).

Generally speaking though, and in line with concerns relating to reliability, statistical learning measures have been shown to only weakly correlate amongst each other (e.g. Schmalz et al., 2019; Siegelman & Frost, 2015), which may help explain the mixed results regarding the relationship between statistical learning and measures of linguistic performance (i.e. some studies reporting significant correlations and others reporting null findings). Of course, these factors do not exclude the possibility that statistical learning plays an important role in language acquisition and is therefore related to children's grammatical performance, but merely affect our ability to evaluate this link (Arnon, 2019a; 2019b). More research is needed in order to improve on present methodologies of measuring statistical learning and to more reliably evaluate its relationship to language.

Finally, we would like to return to lexical storage and/or retrieval as potential additional sources of variation in grammatical performance, and of grammatical difficulties in dyslexia. Of course, lexical knowledge in general is one of the crucial building blocks of the comprehension and production of language, and lexical knowledge is affected in children with DLD (see McGregor, 2009, for a review). This relationship is also apparent from the present study: children's receptive vocabulary knowledge contributes to their performance on inflectional morphology and syntax. More specifically, however, children with dyslexia were shown to experience difficulties in irregular plurals (CELF word structure), irregular past tense (CELF recalling sentences), and the choice between the common and neuter definite article (CELF recalling sentences). We would like to speculate that, besides phonological processing and memory, the automatic access and retrieval of lexical representations may be impaired in dyslexia (see also Bexkens et al., 2014), while the representations themselves may be unimpaired. A similar line of reasoning has been suggested regarding the retrieval processes of representations of speech sounds and phonology (Boets et al., 2013; Griffiths & Snowling, 2001; Ramus & Szenkovits, 2008; Rispens, Baker, & Duinmeijer, 2015). If individuals with dyslexia are unable to efficiently retrieve lexical representations from long-term memory (e.g. irregular plural or past tense forms), they are more likely to apply the regular morphological rule instead (e.g. Pinker, 1999), resulting in overgeneralizations as described in the present study.

In summary, deficits in the area of phonological processing, phonological short-term and working memory, as well as lexical retrieval, are likely to contribute to the linguistic performance of children with dyslexia, not only in the area of literacy skills but also regarding inflectional morphology and syntax. These observations fit with suggestions that multiple cognitive deficits may help explain the range of behavioral difficulties associated with dyslexia and other developmental disorders, as well as the comorbidity between different disorders (e.g. Law, Vandermorsten, Ghesquiére & Wouters 2017; Pennington, 2006). Already in 1999, Wolf and Bowers proposed the double deficit hypothesis: impairments in phonology or rapid automatized naming were assumed to cause dyslexia, with more severe problems when both phonological and rapid automatized naming difficulties were present in a single individual. More research is needed to increase our understanding of the exact nature of the underlying causes of dyslexia and to shed light on the so-called "risk factors" of developing developmental disorders such as dyslexia (Pennington, 2006). Investigations of

multiple sources of variance simultaneously, as attempted in the present study, may shed light on these open questions.

Chapter 7

General discussion

A domain-general learning mechanism – i.e. a statistical learning mechanism – is thought to contribute to the acquisition of spoken language and literacy skills in typical development (e.g. Arciuli & Simpson, 2012; Romberg & Saffran, 2010). Additionally, it has been suggested that impairments in the area of statistical learning may explain the observed language difficulties in developmental disabilities, including dyslexia (e.g. Ullman et al., 2019). Over the past decade, these hypotheses have received growing attention, resulting in divergent findings (i.e. some studies report evidence for the relationship between statistical learning and language, while others yield null results). The current dissertation adds to this body of work by investigating the relationship between statistical learning and performance on language and literacy measures in children with and without dyslexia. It employs three statistical learning tasks that span modalities and structure types and it addresses a range of language domains including spoken language skills (syntax, inflectional morphology) as well as literacy skills (technical reading, spelling). Importantly, the experimental studies reported in the present dissertation control for cognitive measures known to relate to statistical learning, such as sustained attention and short-term and working memory (see e.g. Arciuli, 2017, Frost et al., 2019). Moreover, they take into account other variables known to predict language and literacy skills (vocabulary and phonological skills), and other potential participant-level confounds (e.g. children's age, socio-economic status [SES], gender and non-verbal reasoning ability). In doing so, we aimed to (1) provide valuable new evidence regarding the relationship between individual differences in statistical learning and language performance in children with and without dyslexia, and (2) gain more insight into the nature and the extent of the hypothesized statistical learning deficit in dyslexia. In this final chapter, the main findings from the preceding chapters are recapitulated and discussed in relation to the two main research aims (§7.1 and §7.2 respectively). Subsequently, the chapter presents the theoretical and practical implications of the main findings, and sketches avenues for future research (§7.3). Finally, a number of conclusions are drawn in §7.4.

7.1 Statistical learning in relation to individual differences in language and literacy skills

The relationship between individual differences in statistical learning ability on the one hand and performance on tasks assessing language and literacy skills were investigated in chapters 5 and 6. While chapter 5 focused on the link between variation in serial reaction time (SRT) and visual statistical learning (VSL) performance and technical reading and spelling skills, chapter 6 examined the association of SRT and auditory nonadjacent dependency learning (A-NADL) performance with children's grammar scores. Here, we present the main research findings of chapter 5 (§7.1.1) and chapter 6 (§7.1.2), and discuss these findings in conjunction in §7.1.3, where we focus on what our results may mean for the hypothesized role of statistical learning in the acquisition of language and literacy skills (§7.1.3).

7.1.1 Individual differences in literacy skills

Consistent with the hypothesized relationship between statistical learning and literacy skills, previous studies have demonstrated positive correlations between measures of statistical learning and reading and spelling ability. Further evidence comes from studies reporting poor statistical learning in individuals with dyslexia (see e.g. Lum et al., 2013, for a meta-analysis, and see §7.2 for further discussion of statistical learning in children with dyslexia). Here, we focus on studies of individual differences in statistical learning ability and literacy scores. Arciuli and Simpson (2012) showed that variation in VSL performance relates to variation in word reading abilities, both in English-speaking adult and typically developing (TD) child participants. Since then, others have also reported correlations between statistical learning measures and sentence reading in English (Qi et al., 2019), word reading in Norwegian (von Koss Torkildsen et al., 2019), and reading Hebrew as a second language (Frost et al., 2013) in TD populations. Whereas some of these studies have controlled for age (Arciuli & Simpson, 2012), or age and non-verbal intelligence (Qi et al., 2019), the study by von Koss Torkildsen et al. (2019) was the first to evaluate the relative contribution of statistical learning

while also taking into account other factors known to influence literacy skills (e.g. attention, rapid automatized naming [RAN], verbal working memory).

In chapter 5, we built on these previous studies by looking at two measures of statistical learning (VSL and SRT) and by incorporating literacy skills other than word reading (i.e. non-word reading and spelling). Moreover, we investigated whether statistical learning ability is related to literacy skills over and above cognitive abilities known to relate to statistical learning and/or literacy skills (i.e. non-verbal reasoning, sustained attention, verbal short-term and working memory, and RAN) and other variables at the participant level (i.e. age, SES, and gender). Furthermore, our sample consisted of children with a wide range of literacy abilities, since both TD children and children with dyslexia were included. Finally, we calculated the statistical learning tasks' reliability by looking at their internal reliability and consistency, in view of concerns about the reliability of statistical learning measures when used with children (see Arnon, 2019b, for a review).

The regression model presented in chapter 5 yielded no evidence of a relationship between statistical learning and literacy skills above and beyond the aforementioned participant-level variables. Significant simple correlations were found between SRT performance and both spelling (r = .372, p = .0078) and reading scores (r = .348, p = .013) in children with dyslexia. This (weak) relationship between SRT and literacy skills was no longer significant after controlling for other predictors in the model (e.g. age, non-verbal reasoning, attention, and phonological skills). Thus, these findings underline the importance of considering participant-level variables such as children's scores on attention, non-verbal reasoning and known predictors of reading and spelling (e.g. phonological memory) when investigating the (unique) contribution of statistical learning to (language and) literacy skills. Furthermore, chapter 5 provides no evidence for (or against) the hypothesis that statistical learning ability is related to literacy skills in Dutch-speaking 8 to 11-year-olds, which is in line with other studies reporting null findings (e.g. Nigro et al., 2015; Schmalz et al., 2019; West et al., 2017; 2018).

Previous reports of null findings have led to concerns about the reliability of often used statistical learning measures, especially when used in child participants (Arnon 2019a; 2019b, see e.g. Kidd et al., 2017; Siegelman et al., 2017a; 2017b, for discussions of adult participants). Arnon (2019a) investigated whether three common statistical learning measures, including a VSL similar to

the one reported on here, captured stable individual differences in child participants. Her outcomes were unconvincing: internal consistency as measured using split-half reliability varied between r = -.04 (linguistic auditory statistical learning [ASL]) and r = .59 (VSL), whereas the test–retest reliabilities varied between r = .01 (VSL) and r = .33 (non-linguistic ASL), and all reliability measures were concluded to be "well below psychometric standards" (Arnon, 2019a, p. 7). For this reason, it was important to calculate the reliability of the statistical learning measures used in chapter 5. Although we do not have data regarding the test–retest reliability, results showed that the internal consistency and reliability of the tasks used in the present dissertation approached psychometric standards. In the SRT, the split-half reliability coefficient was r = .71, 95% CI [.58 ... .81], and the VSL split-half reliability coefficients for the 2-AFC and 3-AFC questions were r = .70, 95% CI [.55 ... .80] and r = .78, 95% CI [.67 ... .85] respectively. Therefore, we believe the results presented here are in line with claims that the true effect of statistical learning on the development of literacy skills may only be small, and may consequently only surface under certain methodological conditions (see also Schmalz et al., 2019; Elleman et al., 2019, and §7.1.3 for an elaboration on this mixed pattern of findings in the field, i.e. some studies reporting significant correlations and other studies reporting null results). Nevertheless, the development of novel statistical learning measures that can more reliably assess children's statistical learning ability is an important aim for future research (see §7.3).

While the results regarding the relationship between statistical learning and literacy attainment were inconclusive, chapter 5 provided us with another finding that we would like to reiterate here. Phonological skills, as measured through phonological short-term memory (NWR-S and digit span tasks) and RAN, were found to relate to literacy scores overall, and this effect was larger in participants with dyslexia than in those without (see §7.3 for the theoretical implications for theories of dyslexia). These findings replicate those of earlier work, which has pointed out the relationship between individual differences in phonological skills, such as RAN, and literacy attainment in children (e.g. Furnes & Samuelsson, 2010; Papadopoulos, Spanoudis, & Georgiou, 2016; see Araújo, Reis, Petersson, & Faísca, 2015 for a meta-analysis of 137 studies). Similarly, tasks that assess phonological processing and short-term and working memory, including the NWR-S and digit span tasks used in chapter 5, have been found to relate to literacy scores (e.g. de Bree et al., 2010; de Jong & van der Leij, 1999).

In our study, the effect of phonological skills was largely carried by the effect of the RAN letters subtest, which is supported by the meta-analytical finding that RAN tasks with letters (or numbers) show higher correlations with reading than RAN tasks using pictures (or colors; Araújo et al., 2015). This makes sense, given the fact that reading involves the quick processing, retrieval, and articulation of grapheme–phoneme correspondences, just like the RAN letters (see also the brochure of the Dutch Dyslexia Association [*Stichting Dyslexie Nederland*]; de Jong et al., 2016). Since RAN performance is influenced by a range of underlying processes, including attention, knowledge of phonemes and graphemes, memory, and articulation (see e.g. Papadopoulos et al., 2016), the exact reason why individual differences in RAN relate to literacy scores is unclear (e.g. Kirby, Georgiou, Martinussen, & Parrila, 2010). Nevertheless, our results once again stress the contribution of phonological skills, and RAN letters in particular, to literacy skills in children.

In summary, although chapter 5 did not find evidence of a contribution of statistical learning ability to children's literacy scores above and beyond participant-level variables such as attention, non-verbal reasoning and age, it did replicate earlier findings that stress the important role that phonological skills (especially RAN letters) play in the acquisition of literacy skills in children with and without dyslexia.

7.1.2 Individual differences in grammatical skills

Statistical learning has been claimed to facilitate not only learning to read and spell, but also the development of spoken language. More specifically, statistical learning is hypothesized to support the acquisition of rule-based aspects of language such as morphology and syntax (e.g. Ullman & Pierpont, 2005; Wijnen, 2013). In support of this hypothesis, impairments in statistical learning have been observed in children with developmental language disorder (DLD), who are known to experience difficulties in this area (see e.g. Lammertink et al., 2017; Lum et al., 2014, for meta-analyses). Other support comes from correlational studies: for example, individual differences in statistical learning ability correlate with sentence comprehension in adult speakers (Misyak et al., 2010; Misyak & Christiansen, 2012). Studies with child participants have also provided evidence for an association between statistical learning and grammatical ability (e.g.

passives: Kidd, 2012; object-relative clauses: Kidd & Arciuli, 2016; grammatical processing: Clark & Lum, 2017). A range of other studies have reported non-significant correlations between statistical learning and grammar (e.g. Conti-Ramsden et al., 2015; Gabriel et al., 2011). Since studies vary widely in their methods (e.g. measures of learning, measures of grammar) and in their obtained findings (i.e. some report significant relationships and others report null results), Hamrick, Lum, & Ullman (2018) conducted a meta-analysis of studies investigating the correlation between statistical learning ability, as assessed using the SRT task, and a range of grammatical measures in TD children. Collapsing over eight experimental studies, they report a significant link between SRT performance and grammatical abilities (mean weighted effect = .27, p = .043; but see Lammertink et al., 2019b, for opposing findings in a meta-analysis including children with and without DLD).

Chapter 6 follows up on this line of research by investigating the hypothesized relationship between statistical learning and grammatical performance. It does so elaborately by using two measures of statistical learning (SRT and A-NADL), by looking at both inflectional morphology and syntax as outcome measures, and by including children with and without a diagnosis of dyslexia. Children with dyslexia are an interesting test case, since they may experience (subtle) problems in the area of spoken language (see Snowling & Melby-Lervåg, 2016, for a meta-analysis of children with a familial risk of dyslexia). Different from many previous studies, we controlled for individual differences in cognitive abilities known to be related to statistical learning and/or grammatical performance, including sustained attention, phonological short-term and working memory, and vocabulary.

From our investigation in chapter 6, we could not conclude whether or not statistical learning ability, as measured using the SRT and A-NADL tasks, contributes to children's grammatical performance. The studies by West et al. (2017; 2018), already discussed with respect to literacy skills, also report null correlations between measures of implicit learning (including the SRT task) and a standardized measure of grammar. Moreover, they showed that the implicit learning tasks used in their studies had poor reliability (SRT split-half reliabilities between .17 and .75; West et al., 2017). As briefly mentioned, a recent meta-analysis by Lammertink et al. (2019b) that included a further 11 studies of the relationship between SRT performance and expressive grammatical abilities, found no evidence for a positive relationship overall (r = .13, 95% CI [-.038 –

+.28]). As argued by Lammertink et al. (2019b), these results may be partially explained by the low reliability of statistical learning measures when used with child participants (see also Arnon, 2019a; 2019b). However, the split-half reliabilities of the statistical learning measures used in the present study approached standards for psychometric testing (SRT: $r = .71$, A-NADL: $r = .79$ in a sample of DLD and TD children; see Lammertink et al., 2019a). To conclude, the findings in chapter 6 provide no evidence for (or against) the hypothesis that individual differences in statistical learning ability relate to grammatical skills (see §7.1.3, where we elaborate on possible interpretations of these null findings in relation to the hypothesized role statistical learning plays in the acquisition of language and literacy skills).

As described for chapter 5, the regression model in chapter 6 provided some results that were unrelated to the relationship between statistical learning and grammatical skills that we would nevertheless like to discuss here. We found that phonological processing and phonological short-term and working memory (i.e. the NWR-S and digit span tasks) are related to individual variation in grammatical performance, an effect that was mostly carried by the NWR-S. These findings emphasise the important role that phonological processing and phonological memory may play in the acquisition of grammar (see also e.g. Robertson & Joanisse, 2010). Furthermore, together with the observed relationship between phonological skills and literacy outcomes (van Witteloostuijn et al., under review, see chapter 5), they suggest that the difficulties with written and spoken language experienced by children with dyslexia may be (partially) explained by an underlying problem with phonological skills (e.g. Shankweiler et al., 1995; see §7.3 for implications of the present findings for theories of dyslexia).

To summarize, the findings reported in chapter 6 provide no evidence for (or against) a relationship between statistical learning and grammatical performance in Dutch-speaking children with and without dyslexia. They did, however, stress the important role that phonological processing and phonological memory may play in grammatical performance. We will now present a discussion of the contribution of statistical learning to individual differences in (written) language skills overall, after which we will discuss the results regarding the statistical learning abilities of children with dyslexia (§7.2).

7.1.3 Interim conclusions: individual differences

As we have seen so far, chapters 5 and 6 provide no evidence of the hypothesized role that statistical learning plays in the acquisition of language and literacy skills. Since the start of this PhD project in May 2015, a number of studies have similarly obtained null results regarding the correlation between statistical learning and literacy skills, despite the promising findings reported earlier (e.g. Arciuli & Simpson, 2012; Frost et al., 2013). Examples include studies that examine adult participants (e.g. Schmalz et al., 2019), TD child participants (e.g. West et al., 2017; 2018) and child participants with and without dyslexia (Nigro et al., 2015), and that employ two types of commonly used statistical learning measures: the SRT task (Schmalz et al., 2019; West et al., 2017; 2018) and the AGL task (Nigro et al., 2015; Schmalz et al., 2019).

Recent reviews, as well as the studies in this dissertation, have paid attention to these null findings in the field. Arnon (2019b), for example, highlights a number of general concerns with relating individual differences in performance on statistical learning tasks to variation in (language and) literacy attainment that may help explain spurious results in either direction, the largest concern being the reliability of statistical learning measures to capture individual differences (see e.g. Kidd et al., 2017; Siegelman et al., 2017a; 2017b, for discussions of adults). In our statistical learning experiments, we aimed to increase our tasks' reliabilities by including online measures (A-NADL; Siegelman et al., 2017b) and by using different types of offline questions (VSL; Siegelman et al., 2017a), which resulted in relatively high internal consistency and reliability within sessions, with split-half reliabilities varying between .70 (VSL 2-AFC) and .79 (A-NADL). Note, however, that we have no data on test–retest reliability, which is a measure of reliability within individuals, measured between sessions (i.e. the correlation between participants' individual performance on two instances of a single measure, usually administered a few weeks or months apart). As such, test–retest reliability informs us whether a certain measure captures a stable trait of an individual and would provide additional evidence that the statistical learning tasks used in the present dissertation can be described as reliable measures (Arnon, 2019a). Developing both on- and offline measures of statistical learning that can reliably assess individual differences remains a current challenge, especially for child participants. The use of (relatively) unreliable

measures may help explain the observed fluctuations between studies correlating statistical learning performance with language and literacy attainment (see also §7.3).

An additional explanation for the mixed pattern of findings in studies investigating the relationship between statistical learning ability and language and literacy skills is that the true effect may be small (or zero) and difficult to detect. If that is the case, the effect may only be observable under certain methodological conditions, which is supported by the suggestion that methodological differences between studies may help explain why some studies find evidence of the relationship between statistical learning and linguistic performance, while others do not (see e.g. Elleman et al., 2019; Schmalz et al., 2019). As indicated in chapter 5, more specific theoretical and pedagogical models of language and literacy acquisition should be formulated with regard to statistical learning. Such models should focus on the specific role that statistical learning plays in the process of learning to speak, read and spell, and could be guided by questions such as "when during development is statistical learning most important?" and "what type of statistical structure is most closely related to the structures observed in spoken and written language?". On the basis of this type of models, researchers may formulate testable hypotheses that, for example, guide the choice of certain statistical learning measures when looking at different components of language (see also e.g. Frost et al., 2019; Lammertink et al., 2019b; Siegelman et al., 2017b). In doing so, future research may more closely target the methodological conditions that may inform us about the nature and the extent of the correlation between statistical learning performance and (language and) literacy skills.

Of course, these concerns do not deny the idea that a human capacity for learning statistical structures is an important contributor to the acquisition of language and literacy skills, but they do hamper our ability to draw conclusions at this point (see also Arnon, 2019b). Moreover, they provide possible avenues for future research, which we will get back to in §7.4.

7.2 Statistical learning in dyslexia

Besides studies of individual differences, the present dissertation contains two chapters that focus on the statistical learning abilities of children with dyslexia as compared to those without. In chapter 3, we reported a meta-analysis of 13

previous studies that tested the statistical learning deficit in dyslexia through the visual AGL paradigm. In the subsequent chapter, we investigated potential group differences (i.e. dyslexia versus control) on three statistical learning tasks that varied across domains and the type of statistical structure targeted. Through this elaborate investigation of statistical learning in dyslexia, both through meta-analytical and experimental techniques, we hope to gain insight into the hypothesized role that statistical learning plays in the reading and spelling problems that are observed in children with dyslexia. In §7.2.1, we first present the meta-analytical findings, followed by the experimental findings in §7.2.2, after which these results are discussed together in §7.2.3.

7.2.1 Meta-analytical findings: statistical learning in dyslexia

As is the case for other experimental paradigms, studies of the visual AGL performance of individuals with dyslexia have yielded a mixed pattern of results: while some studies report significant differences between participants with and without dyslexia (e.g. Ise et al., 2012; Laasonen et al., 2014), other studies report null results (e.g. Nigro et al., 2016; Rüsseler et al., 2006). In 2017, Schmalz et al. performed a systematic review of both SRT and AGL studies in dyslexia, and concluded that there is "insufficient high-quality data to draw conclusions about the presence or absence of an effect" (Schmalz et al., 2017, p.147). We elaborated on their review in chapter 3 by (a) including a larger set of (unpublished) studies, (b) adding a statistical examination of the possibility of a publication bias in the field, and (c) investigating the effect of methodological variables (e.g. age, stimulus type, and training method) through a meta-regression technique, focusing on the visual AGL paradigm.

Although the overall average weighted effect size obtained in chapter 3 was significant (0.46, 95% CI [0.14 … 0.77], $p = 0.008$), indicating poorer visual AGL performance in individuals with dyslexia as compared to control participants across 13 studies, these results should be interpreted with caution. As reported in §3.3.2, there were indications of a publication bias. When we (conservatively) corrected for this publication bias, the estimated effect size was considerably reduced and the effect of group on performance no longer reached significance (0.20, 95% CI [−0.11 … 0.50], $p = 0.205$). This means that the initial results from our meta-analysis are likely to be overly optimistic, and the effect

may well be nulled by unpublished findings that were not included in the analysis. These findings highlight the importance of more empirical studies on visual AGL performance specifically, and statistical learning abilities more generally, in comparisons of individuals with and without dyslexia, before a conclusion can be drawn regarding the hypothesized deficit in this area (see also Schmalz et al., 2017; Van Elk, Matzke, Gronau, Guan, Vandekerckhove, & Wagenmakers, 2015). Furthermore, the publication bias that we likely observed reflects the known problem in psychological research that significant findings are more likely to be published than non-significant ones, which may impact both scientific and public perception of research outcomes (see also e.g. Ferguson & Brannick, 2012; Rosenthal, 1979).

7.2.2 Experimental findings: statistical learning in dyslexia

In an attempt to gain further insight into the hypothesized role that statistical learning ability may play in explaining the observed reading and spelling difficulties in dyslexia, chapter 4 investigated the performance of school-aged children with dyslexia on three statistical learning paradigms other than the visual ALG task. As mentioned, we included not only the often-used visuo-motoric SRT task, but also introduced two relatively novel tasks: the self-paced VSL task and an A-NADL task that also makes use of an online RT measure of learning. Additionally, investigations of group differences were controlled for individual differences in sustained attention and verbal short-term memory. This was done because (a) sustained attention and short-term memory capacity have been argued to play an important role in performance on statistical learning tasks (e.g. Arciuli, 2017; Arciuli & Simpson, 2011; Baker et al., 2004; Toro et al., 2005), and (b) children with dyslexia were found to perform worse on these cognitive constructs (see §4.2.1).

The results revealed significant learning effects overall (i.e. collapsing over participants with and without dyslexia) on the online RT measures of the SRT and A-NADL tasks and on the offline accuracy measure of the VSL task. However, no significant learning was found for the VSL online RT measure and the A-NADL offline accuracy measure, indicating the difficulty of assessing the statistical learning abilities of school-aged children (see also e.g. Arnon 2019a; 2019b; van Witteloostuijn, Lammertink et al., 2019, see chapter 2, and see §7.2.3

below for more detail). More importantly, no significant differences between children with and without dyslexia were found for any of the on- or offline measures of statistical learning. Thus, the experimental findings reported in chapter 4 are consistent: although we find evidence that suggests sensitivity to the statistical structures in the SRT, VSL and A-NADL tasks by school-aged children overall, we find no evidence for (or against) a statistical learning deficit in children with dyslexia.

7.2.3 Interim conclusions: group effects

Despite the fact that the findings presented in chapters 3 and 4 are not in line with the hypothesized domain-general statistical learning deficit in dyslexia, they are consistent: we find no evidence that individuals with dyslexia have difficulties with statistical learning measures across our experimental study of three paradigms (SRT, VSL, A-NADL), and we find no meta-analytical evidence of difficulties with visual AGL once we controlled for the presence of a publication bias. Rather, we find evidence of sensitivity to the statistical structures in the SRT, VSL, and A-NADL tasks when we look at the group of children with and without dyslexia combined, and no evidence of a difference between the two groups. Thus, given the fact that (a) participants showed evidence of learning in all three experimental tasks in chapter 3, and (b) the standardized effect size of the group effect was small for all three experimental tasks in chapter 3, as well as on the (for publication bias corrected) meta-analytical results of the visual AGL in chapter 4, and (c) the split-half reliabilities of the three experimental tasks were found to be relatively high in chapters 5 and 6, the results of the present dissertation do not support the hypothesis that individuals with dyslexia exhibit a domain-general statistical learning deficit. Since we did not provide evidence of such a deficit, this raises doubt about the possibility that a statistical learning deficit causes the literacy problems seen in individuals with dyslexia. The implications of these results for theories of dyslexia are discussed in §7.3.

These findings are in line with other studies that have yielded null results (e.g. Deroost et al., 2010; Menghini et al., 2010; Nigro et al., 2016; Rüsseler et al., 2006; Staels & Van den Broek, 2017). They also support concerns relating to a publication bias in the field (e.g. Schmalz et al., 2017; van Witteloostuijn et al., 2017, see chapter 3), and studies that have questioned (the extent and nature of)

the statistical, or procedural, learning deficit hypothesis of dyslexia (see also e.g. West et al., 2017; 2018). As discussed in relation to studies of individual differences, the results of the present dissertation do not exclude subtler, or perhaps more domain-specific, statistical learning difficulties in dyslexia that may only surface under certain experimental conditions (e.g. statistical learning of certain types of structure or in certain modalities or stimulus types that relate more closely to literacy skills). Furthermore, generally speaking, the problem of the reliability of statistical learning tasks may also impact the field of statistical learning in dyslexia.

Besides informing us about the statistical learning abilities of children with dyslexia, the results in chapter 4 demonstrate the difficulty of measuring the statistical learning abilities of school-aged children. This topic already received attention in chapter 2 (van Witteloostuijn, Lammertink et al., 2019), where we showed that our online, but not offline, measure of learning was able to capture VSL ability in children between 5;9 and 8;7 years old (note that the effect of online learning was rather weak with a *p*-value of .021). Together with the results from chapter 4 – i.e. no evidence of learning was observed in the VSL online measure and the A-NADL offline measure – the findings in the present dissertation underline the importance of developing sensitive measures of statistical learning in school-aged children, despite the relatively high split-half reliabilities of the measures that *did* show evidence of learning (VSL offline measures, SRT online measure, and A-NADL online measure).

The fact that the VSL offline measures did not provide evidence of learning in 5- to 8-year-old children (chapter 2), but did provide evidence of learning in 8- to 11-year-old children (chapter 4), is indirect evidence that performance on offline, explicit decision-making measures increase between the ages of 5 and 12 (Arciuli & Simpson, 2011; Arnon & Raviv, 2017). Since the CIs of the online effect of learning in the self-paced VSL task (i.e. the difference in RTs to predictable and unpredictable stimuli) in the two samples overlap (chapter 2: CI = [-0.114 … -0.002]; chapter 4: CI = [-0.038 … +0.012]), we can conclude on the basis of these results that the online measure does not (yet) reliably measure children's sensitivity to the statistical structure of the VSL task. Likewise, the offline measure of the A-NADL was shown to be insensitive to children's ability to track the nonadjacent dependencies in the task (see also Lammertink et al., 2019a; Lammertink, van Witteloostuijn et al., 2019). These results stress the need for additional sensitive measures of statistical learning in child participants,

although they have also demonstrated the usefulness of the SRT, the VSL offline and the A-NADL online measures adopted in the present dissertation.

7.3 Implications and directions for future research

As stated at the beginning of this chapter, a domain-general statistical learning mechanism has been hypothesized to play a crucial role in the acquisition of spoken and written language. This hypothesis has been assessed in two traditional types of research: (1) the study of individual differences in statistical learning and their relationship with variation in language and literacy scores, and (2) the study of the statistical learning performance of children with developmental disorders such as dyslexia. These two approaches were combined in the present dissertation. Taken together, the findings obtained have led us to question the strength of this relationship between statistical learning on the one hand and language and literacy skills (and dyslexia) on the other hand. As previously stated, however, our findings do not exclude the possibility that such a relationship exists, although the association may be subtler and perhaps more domain-specific than previously hypothesized. Here, we wish to discuss some implications and highlight some directions for future research, before turning to the conclusions in §7.4.

The first important implication of the studies presented in this dissertation is practical in nature. We have shown that the three statistical learning measures used in the present dissertation are sensitive to the statistical learning abilities of school-aged children. Particularly, the VSL offline test phase, which was improved based on suggestions made by Siegelman et al. (2017b), and the A-NADL online RT measure as modeled on a previous study with adults by López-Barroso (2016, see also Lammertink, van Witteloostuijn et al., 2019) were novel measures that were found to capture learning on the group level. Moreover, the split-half reliabilities of the measures used in this dissertation approached psychometric standards, in contrast to earlier reports of low reliabilities of statistical learning measures in children (e.g. West et al., 2017; Arnon, 2019a). Thus, these measures are suitable methods for future studies investigating the statistical learning abilities of school-aged children, perhaps in relation to other cognitive abilities.

Nevertheless, the investigations presented in this dissertation also reveal some difficulties with assessing statistical learning in children, and highlight the need for statistical learning measures that are more sensitive to learning in child populations. Specifically, future studies should aim to improve current online measures of learning in the VSL task, as well as offline measures of learning in case of the A-NADL task, since these measures did not provide evidence of learning (van Witteloostuijn et al., 2019, see chapter 4). Moreover, the reliability of statistical learning measures for assessing individual differences could be further improved, and investigated extensively through test–retest reliability in addition to split-half reliability as reported here. Improved measures of statistical learning may help clarify the true relationship between statistical learning on the one hand, and language and literacy acquisition on the other hand. Moreover, they may better inform us about the nature and the extent of the (potential) statistical learning problems in language-based developmental disorders, including dyslexia.

Besides stressing the need for the development of more sensitive measures of statistical learning, the mixed pattern of findings in the field (i.e. some studies reporting significant findings and other studies reporting null results) calls for replications and large-scale studies. Further, the use of pre-registered reports may help minimize some problems in the field, including the publication bias as signaled in van Witteloostuijn et al. (2017, see chapter 3). With the accumulation of increasingly reliable evidence, the use of meta-analyses may eventually allow us to draw conclusions regarding (1) the relationship between statistical learning ability and performance on measures of language and literacy skills, and (2) the statistical learning abilities of individuals with dyslexia.

Regarding the language and literacy skills of children with dyslexia, the studies in the present dissertation have found deficits not only in the expected areas of technical reading, spelling, phonological skills (i.e. phonological processing, phonological short-term and working memory) and lexical retrieval (i.e. RAN), but also more subtle deficits in inflectional morphology and syntax (van Witteloostuijn et al., submitted, see chapter 6). These grammatical problems were evidenced by errors with retrieving the correct (irregular) plural and (irregular) past tense from the lexicon (e.g. plural: *ei*, *ei–eren*; 'egg–PL'; e.g. past tense: *koop–*te*; 'buy–*ed'), errors concerning the correct definite article choice (i.e. common *de* or neuter *het*) and errors regarding word order (i.e. verb second:

**zij afwassen*, *'they washing up'). These (exploratory) error pattern findings merit future research into the grammatical performance of individuals with dyslexia.

Scores on phonological processing, phonological short-term and working memory, and lexical retrieval were shown to relate to children's performance both on tasks that assess literacy skills (van Witteloostuijn et al., under review, see chapter 5) and on tasks that assess grammatical skills (van Witteloostuijn et al., submitted, see chapter 6). Thus, future research that intends to explore the unique contribution of statistical learning to language and literacy acquisition, should control for these constructs that are known to predict children's variability in language and literacy skills (e.g. phonological skills, RAN, vocabulary, non-verbal IQ, sustained attention; see also von Koss Torkildsen et al., 2019). Similarly, future investigations of the statistical learning abilities of individuals with dyslexia should consider these constructs, since dyslexia is associated with problems in the areas of phonological skills, attention and memory. From the present dissertation, it appears likely that dyslexia is associated with multiple cognitive deficits, including impairments in phonology and lexical retrieval (see e.g. Law, Vandermorsten, Ghesquiére & Wouters 2017; Pennington, 2006; Wolf & Bowers, 1999). More research is needed to increase our understanding of the underlying causes of dyslexia, and to point out "risk factors" of developing dyslexia (Pennington, 2006). Future studies that investigate the underlying deficits in dyslexia should therefore consider multiple potential cognitive deficits simultaneously, as was attempted here, to clarify their relative contributions to (language and) literacy problems.

7.4 Conclusions

The introduction of this dissertation raised the question "why are children such efficient language learners?", and posited the hypothesis that a domain-general learning mechanism – i.e. statistical learning – could explain children's implicit inference of abstract patterns and rules in written and spoken language. What we have shown in chapters 2 and 4, is that children indeed have the capacity to pick up statistical structures presented to them across domains (i.e. visuo-motoric, visual, and auditory) and structure types (i.e. adjacent and nonadjacent relationships), and the statistical learning measures used in the present study were able to measure this sensitivity. However, chapters 5 and 6 provide no evidence

of a direct relationship between performance on these statistical learning tasks and children's scores on tests of reading, spelling, inflectional morphology and syntax. Furthermore, chapters 3 and 4 did not provide (convincing, in the case of our meta-analytical results in chapter 3) evidence of a domain-general statistical learning deficit in individuals with dyslexia. In conclusion, it cannot be excluded that the link between statistical learning ability and language and literacy acquisition may be less strong than hypothesized and may only surface under certain methodological conditions (see also e.g. Schmalz et al., 2019; Elleman et al., 2019. These findings, unfortunately, do not bring us closer to unraveling the underlying cause of dyslexia. Rather, it appears likely that individuals with dyslexia do not experience domain-general, extensive problems with statistical learning. Moving forward, large-scale and pre-registered studies, as well as meta-analyses, are needed to allow us to reach conclusions regarding the contribution of (domain-general) statistical learning ability to the acquisition of language and literacy skills, both in typical and in impaired populations.

References

Abrams, M., & Reber, A.S. (1988). Implicit learning: Robustness in the face of psychiatric disorders. *Journal of Psycholinguistic Research, 17*(5), 425–439.

Alloway, T.P. (2012). *Alloway Working Memory Assessment (AWMA)*. London, England: Pearson.

Alloway, T.P., Gathercole, S.E., Kirkwood, H., & Elliott, J. (2009). The cognitive and behavioral characteristics of children with low working memory. *Child Development, 80*(2), 606–621.

Apfelbaum, K.S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading: Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology, 49*(7), 1348–1365.

Araújo, S., Reis, A., Petersson, K.M., & Faísca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *Journal of Educational Psychology*, *107*(3), 868–883.

Archibald, L.M., & Gathercole, S.E. (2006). Short-term and working memory in specific language impairment. *International Journal of Language & Communication Disorders, 41*(6), 675–693.

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), Article 20160058.

Arciuli, J. (2018). Reading as statistical learning. *Language, Speech, and Hearing Services in Schools, 49*(3S), 634–643.

Arciuli, J., & Conway, C.M. (2018). The promise—and challenge—of statistical learning for elucidating atypical language development. *Current Directions in Psychological Science, 27*(6), 492–500.

Arciuli, J., & Cupples, L. (2006). The processing of lexical stress during visual word recognition: Typicality effects and orthographic correlates. *Quarterly Journal of Experimental Psychology, 59*(5), 920–948.

Arciuli, J., & Simpson, I.C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science, 14*(3), 464–473.

Arciuli, J., & Simpson, I.C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286–304.

Arnon, I. (2019a). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, Advance online publication: DOI: 10.10.3758/s13428-019-01205-5

Arnon, I. (2019b). Statistical learning, implicit learning, and first language acquisition: A critical evaluation of two developmental predictions. *Topics in Cognitive Science, 11*(3), 504–519.

Aro, T., Eklund, K., Nurmi, J.E., & Poikkeus, A.M. (2012). Early language and behavioral regulation skills as predictors of social outcomes. *Journal of Speech, Language, and Hearing Research, 55*(2), 395–408.

Aslin, R.N., & Newport, E.L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning, 64*(S2), 86–105.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29.

Baguley T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Houndmills, England: Palgrave Macmillan.

Baker, C.I., Olson, C.R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15*(7), 460–466.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences, 106*(41), 17284–17289.

Bar-Shalom, E.G., Crain, S., & Shankweiler, D. (1993). A comparison of comprehension and production abilities of good and poor readers. *Applied Psycholinguistics, 14*(2), 197–227.

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *Package "lme4": Linear mixed-effects models using Eigen and S4* [R package version 1.1–7], http://CRAN.R-project.org/package=lme4

Batterink, L.J., & Paller, K.A. (2017). Online neural monitoring of statistical learning. *Cortex, 90*, 31–45.

Bertels, J., Franco, A., & Destrebecqz, A. (2012). How implicit is visual statistical learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1425–1431.

Bertels, J., Boursain, E., Destrebecqz, A., & Gaillard, V. (2015). Visual statistical learning in children and young adults: How implicit? *Frontiers in Psychology, 5*, Article 1541.

Bexkens, A., van den Wildenberg, W.P., & Tijms, J. (2015). Rapid automatized naming in children with dyslexia: Is inhibitory control involved? *Dyslexia, 21*(3), 212–234.

Bialystok, E. (1986). Factors in the growth of linguistic awareness. *Child Development, 57*(2), 498–510.

Bishop, D.V.M., & Snowling, M.J. (2004). Developmental dyslexia and specific language impairment: Same or different? *Psychological Bulletin, 130*(6), 858–886.

Bishop, D.V.M., Snowling, M.J., Thompson, P.A., Greenhalgh, T. (2017). CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development. *Journal of Child Psychology & Psychiatry, 58*(10), 1068–1080.

Blom, E., Polišenská, D., & Weerman, F. (2008). Articles, adjectives and age of onset: The acquisition of Dutch grammatical gender. *Second Language Research, 24*(3), 297–331.

Blom, E., Vasić, N., & de Jong, J. (2014). Production and processing of subject–verb agreement in monolingual Dutch children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 57*(3), 952–965.

Boersma, T.A. (2018) *Variability in the acquisition of alomorphs: The Dutch diminutive and past tense* (Doctoral dissertation). University of Amsterdam, Amsterdam, The Netherlands.

Boets, B., Op de Beeck, H., Vandermosten, M., Scott, S.K., Gillebert, C.R., Mantini, D., Bulthé, J., Sunaert, S., Wouters, J. & Ghesquière, P. (2013). Intact but less accessible phonetic representations in adults with dyslexia. *Science, 342*(6163), 1251–1254.

Bond Jr, C.F., Wiitala, W.L., & Richard, F.D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*(4), 406–418.

Borenstein, M.H., Higgins, L.V., & Rothstein, J.P.T. (2009). *Introduction to meta-analysis*. Hoboken, United States: John Wiley & Sons.

Bosse, M.L., Tainturier, M.J., & Valdois, S. (2007). Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition, 104*(2), 198–230.

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

Brus B., & Voeten M. (1972) *Eén Minuut Test* [Measurement instrument]. Nijmegen, The Netherlands: Berkhout Testmateriaal.

Buchholz, J., & Davies, A.A. (2005). Adults with dyslexia demonstrate space-based and object-based covert attention deficits: Shifting attention to the periphery and shifting attention between objects in the left visual field. *Brain and Cognition, 57*(1), 30–34.

Bull, R., Espy, K.A., & Wiebe, S.A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology, 33*(3), 205–228.

Bussy, G., Krifi-Papoz, S., Vieville, L., Frenay, C., Curie, A., Rousselle, C., Rougeot, C., Des Portes, V., & Herbillon, V. (2011). Apprentissage procédural implicite dans la dyslexie de surface et la dyslexie phonologique. *Revue de Neuropsychologie, 3*(3), 141–146.

Capel, D. (2018). *Sequential learning, domain generality, and developmental dyslexia* (Doctoral dissertation). Utrecht University, Utrecht, The Netherlands.

Carroll, J.M., & Myers, J.M. (2010). Speech and language difficulties in children with and without a family history of dyslexia. *Scientific Studies of Reading, 14*(3), 247–265.

Catts, H.W., Adlof, S.M., Hogan, T.P., & Weismer, S.E. (2005). Are specific language impairment and dyslexia distinct disorders? *Journal of Speech, Language, and Hearing Research, 48*(6), 1378–1396.

Chen, A., Wijnen, F., Koster, C., & Schnack, H. (2017). Individualized early prediction of familial risk of dyslexia: A study of infant vocabulary development. *Frontiers in Psychology, 8*, Article 156.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York, United States: Praeger.

Christiansen, M.H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science, 11*(3), 468–481.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (second edition, p. 567). Hillsdale, United States: Erlbaum.

Cohen, A., Ivry, R.I., and Keele, S.W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 17–30.

Conti-Ramsden, G., Durkin, K., Toseeb, U., Botting, N., & Pickles, A. (2018). Education and employment outcomes of young adults with a history of developmental language disorder. *International Journal of Language & Communication Disorders, 53*(2), 237–255.

Conway, C.M., Pisoni, D.B., Anaya, E.M., Karpicke, J., & Henning, S.C. (2011). Implicit sequence learning in deaf children with cochlear implants. *Developmental Science, 14*(1), 69–82.

Copas, J. (1999). What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 162*(1), 95–109.

Copas, J., & Shi, J.Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics, 1*(3), 247–62.

Cowan, N., Hogan, T.P., Alt, M., Green, S., Cabbage, K.L., Brinkley, S., & Gray, S. (2017). Short-term memory in childhood dyslexia: Deficient serial order in multiple modalities. *Dyslexia, 23*(3), 209–233.

Daltrozzo, J., & Conway, C.M. (2014). Neurocognitive mechanisms of statistical-sequential learning: What do event-related potentials tell us? *Frontiers in Human Neuroscience, 8*, Article 437.

de Bree, E.H. (2007). *Dyslexia and phonology: A study of the phonological abilities of Dutch children at-risk of dyslexia* (Doctoral dissertation). Utrecht University, Utrecht, The Netherlands.

de Bree, E.H., & Kerkhoff, A. (2010). Bempen or bemben: Differences between children at-risk of dyslexia and children with SLI on a morpho-phonological task. *Scientific Studies of Reading, 14*(1), 85–109.

de Bree, E.H., van der Ven, S., & van der Maas, H. (2017). The voice of Holland: Allograph production in written Dutch past tense inflection. *Language Learning and Development, 13*(3), 215–240.

de Bree, E.H., Wijnen, F.N.K., & Gerrits, E. (2010). Non-word repetition and literacy in Dutch children at-risk of dyslexia and children with SLI: Results of the follow-up study. *Dyslexia, 16*(1), 36–44.

de Jong, J. (1999). *Specific language impairment in Dutch: Inflectional morphology and argument structure* (Doctoral dissertation). University of Groningen, Groningen, The Netherlands.

de Jong, P.F., de Bree, E., Henneman, K., Kleijnen, R., Loykens, E.H.M., Rolak, M., Struiksma, A.J.C., Verhoeven, L., & Wijnen, F.N.K. (2016). *Dyslexie: Diagnostiek en behandeling*. Stichting Dyslexie Nederland.

de Jong, P.F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology, 91*(3), 450–476.

Del Re, A.C. (2014). *Package "compute.es": Compute effect sizes* [R package version 0.2], http://CRAN.R-project.org/package=compute.es

Deroost, N., Zeischka, P., Coomans, D., Bouazza, S., Depessemier, P., & Soetens, E. (2010). Intact first- and second-order implicit sequence learning in secondary-school-aged children with developmental dyslexia. *Journal of Clinical and Experimental Neuropsychology, 32*(6), 561–572.

Destrebecqz, A., and Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review, 8*(2), 343–350.

DSM-IV (2000). *Diagnostic and Statistical Manual of Mental Disorders*: Fourth Edition, Text Revision (DSM-IV-TR). American Psychiatric Association.

DSM-V (2013). *Diagnostic and Statistical Manual of Mental Disorders*: Fifth Edition. American Psychiatric Association.

Du, W. (2013). *Associative Implicit Learning in Adult Dyslexic Readers* (Doctoral dissertation), University of Strathclyde, Glasgow, United Kingdom.

Du, W., & Kelly, S.W. (2013). Implicit sequence learning in dyslexia: A within-sequence comparison of first- and higher-order information. *Annals of Dyslexia, 63*(2), 154–170.

Dunn, L.M., Dunn, L.M. (2005). *Peabody Picture Vocabulary Test*, Third edition in Dutch [Measurement instrument; Dutch version by L. Schlichting]. Amsterdam, The Netherlands: Harcourt Test Publishers.

Duval, S. (2005). The trim and fill method. In H.R. Rothstein, A.J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, England: John Wiley & Sons.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

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

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

Evans, J.L., Saffran, J.R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 52*(2), 321–335.

Facoetti, A., Paganoni, P., & Lorusso, M.L. (2000). The spatial distribution of visual attention in developmental dyslexia. *Experimental Brain Research, 132*(4), 531–538.

Fawcett, A.J., & Nicolson, R.I. (1995). Persistent deficits in motor skill of children with dyslexia. *Journal of Motor Behavior, 27*(3), 235–240.

Fawcett, A.J., & Nicolson, R.I. (2019). Development of dyslexia: The delayed neural commitment framework. *Frontiers in Behavioral Neuroscience, 13*, Article 112.

Ferguson, C.J., & Brannick, M.T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120–128.

Fiser, J., & Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99*(24), 15822–15826.

Folia, V., Uddén, J., Forkstam, C., Ingvar, M., Hagoort, P., & Petersson, K.M. (2008). Implicit learning and dyslexia. *Annals of the New York Academy of Sciences, 1145*(1), 132–150.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ... & Heiberger, R. (2012). *Package "car": Companion to applied regression* [R package version 2.1-5], http://CRAN.R-project.org/package=car

Franco, A., Gaillard, V., Cleeremans, A., & Destrebecqz, A. (2015). Assessing segmentation processes by click detection: Online measure of statistical learning, or simple interference? *Behavior Research Methods, 47*(4), 1393–1403.

Frost, R., Armstrong, B.C., & Christiansen, M.H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153.

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

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science, 24*(7), 1243–1252.

Froyen, D., Willems, G., & Blomert, L. (2011). Evidence for a specific cross-modal association deficit in dyslexia: An electrophysiological study of letter–speech sound processing. *Developmental Science, 14*(4), 635–648.

Furnes, B., & Samuelsson, S. (2010). Predicting reading and spelling difficulties in transparent and opaque orthographies: A comparison between Scandinavian and US/Australian children. *Dyslexia, 16*(2), 119–142.

Gabay, Y., Thiessen, E.D., & Holt, L.L. (2015). Impaired statistical learning in developmental dyslexia. *Journal of Speech, Language, and Hearing Research, 58*(3), 934–945.

Gabay, Y., Schiff, R., & Vakil, E. (2012). Dissociation between the procedural learning of letter names and motor sequences in developmental dyslexia. *Neuropsychologica, 50*(10), 2435–2441.

Garon, N., Bryson, S.E., & Smith, I.M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin, 134*(1), 31–60.

Gathercole, S.E., Alloway, T.P., Willis, C., Adams, A.M. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology, 93*(3), 265–281.

Gathercole, S.E., & Baddeley, A.D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*(3), 336–360.

Gómez, R.L. (1997). Transfer and Complexity in Artificial Grammar Learning. *Cognitive Psychology, 33*(2), 154–207.

Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436.

Gómez, D.M., Bion, R.A., & Mehler, J. (2011). The word segmentation process as revealed by click detection. *Language and Cognitive Processes, 26*(2), 212–223.

Grama, I.C., Kerkhoff, A., & Wijnen, F.N.K. (2016). Gleaning structure from sound: The role of prosodic contrast in learning non-adjacent dependencies. *Journal of Psycholinguistic Research, 45*(6), 1427–1449.

Griffiths, Y.M., & Snowling, M.J. (2001). Auditory word identification and phonological skills in dyslexic and average readers. *Applied Psycholinguistics, 22*(3), 419–439.

Hamrick, P., Lum, J.A., & Ullman, M.T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences, 115*(7), 1487–1492.

Hardy, R.J., & Thompson, S.G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17*(8), 841–856.

He, X., & Tong, S.X. (2017). Quantity Matters: Children with dyslexia are impaired in a small, but not large, number of exposures during implicit repeated sequence learning. *American Journal of Speech-Language Pathology, 26*(4), 1080–1091.

Hedenius, M., Persson, J., Alm, P.A., Ullman, M.T., Howard Jr, J.H., Howard, D.V., & Jennische, M. (2013). Impaired implicit sequence learning in children with developmental dyslexia. *Research in Developmental Disabilities, 34*(11), 3924–3935.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, United States: Academic Press.

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

Hill, E.L. (2001). Non-specific nature of specific language impairment: A review of the literature with regard to concomitant motor impairments. *International Journal of Language & Communication disorders, 36*(2), 149–171.

Howard Jr, J.H., Howard, D.V., Japikse, K.C., & Eden, G.F. (2006). Dyslexics are impaired on implicit higher-order sequence learning, but not on implicit spatial context learning. *Neuropsychologia, 44*(7), 1131–1144.

Humphrey, N., & Mullins, P.M. (2002). Research section: Personal constructs and attribution for academic success and failure in dyslexia. *British Journal of Special Education, 29*(4), 196–203.

Hung, Y.-H., Frost, S.J., Molfese, P. Malins, J.G., Landi, N., Einar Mencl, W., Rueckl, J.G., Bogaerts, L., & Pugh, K.R. (2019). Common neural basis of motor sequence learning and word recognition and its relation with individual differences in reading skill. *Scientific Studies of Reading, 23*(1), 89–100.

Iao, L.S., Ng, L.Y., Wong, A.M.Y., & Lee O.T. (2017). Nonadjacent dependency learning in Cantonese-speaking children with and without a history of specific language impairment. *Journal of Speech, Language, and Hearing Research, 60*(3), 694–700.

Inácio, F., Faísca, L., Forkstam, C., Araújo, S., Bramão, I., Reis, A., & Petersson, K.M. (2018). Implicit sequence learning is preserved in dyslexic children. *Annals of Dyslexia, 68*(1), 1–14.

Ise, E., Arnoldi, C.J., Bartling, J., & Schulte-Körne, G. (2012). Implicit learning in children with spelling disability: Evidence from artificial grammar learning. *Journal of Neural Transmission, 119*(9), 999–1010.

Janacsek, K., Fiser, J., & Nemeth, D. (2013). The best time to acquire new skills: age-related differences in implicit sequence learning across the human life span. *Developmental Science, 15*(4), 496–505.

Janacsek, K., & Nemeth, D. (2012). Predicting the future: From implicit learning to consolidation. *International Journal of Psychophysiology, 83*(2), 213–221.

Janacsek, K., & Nemeth, D. (2015). The puzzle is complicated: When should working memory be related to implicit sequence learning, and when should it not? (Response to Martini et al.). *Cortex*, *64*, 411–412.

Jiménez-Fernández, G., Vaquero, J.M.M., Jiménez, L., & Defior, S. (2011). Dyslexic children show deficits in implicit sequence learning, but not in explicit sequence learning or contextual cueing. *Annals of Dyslexia, 61*(1), 85–110.

Joanisse, M.F., Manis, F.R., Keating, P., & Seidenberg, M.S. (2000). Language deficits in dyslexic children: Speech perception, phonology, and morphology. *Journal of Experimental Child Psychology, 77*(1), 30–60.

Joanisse, M.F., & Seidenberg, M.S. (1998). Specific language impairment: A deficit in grammar or processing? *Trends in Cognitive Sciences, 2*(7), 240–247.

Joanisse, M.F., & Seidenberg, M.S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, 96*(13), 7592–7597.

Jost, E., Conway, C.M., Purdy, J.D., & Hendricks, M.A. (2011). Neurophysiological correlates of visual statistical learning in adults and children. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.)*, Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Kahta, S., & Schiff, R. (2016). Implicit learning deficits among adults with developmental dyslexia. *Annals of Dyslexia, 66*(2), 235–250.

Karuza, E.A., Farmer, T.A., Fine, A.B., Smith, F.X., & Jaeger, T.F. (2014). On-line measures of prediction in a self-paced statistical learning task. *Proceedings of the Cognitive Science Society, 36*(36), 725–730.

Kelly, S. W., Griffiths, S., & Frith, U. (2002). Evidence for implicit sequence learning in dyslexia. *Dyslexia, 8*(1), 43–52.

Kerkhoff, A., de Bree, E.H., de Klerk, M., & Wijnen, F.N.K. (2013). Non-adjacent dependency learning in infants at familial risk of dyslexia. *Journal of Child Language, 40*(1), 11–28.

Kerkhoff, A., de Bree, E.H., & Wijnen, F.N.K. (2017). Can poor readers be good learners? In E. Segers & P. van den Broek (Eds.), *Developmental Perspectives in Written Language and Literacy*. Amsterdam, The Netherlands: John Benjamins.

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

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

Kidd, E., Donnelly, S., & Christiansen, M.H. (2017). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences, 22*(2), 154–169.

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

Kirby, J.R., Georgiou, G.K., Martinussen, R., & Parrila, R. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly, 45*(3), 341–362.

Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), B35–B42.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*(17), 2693–2710.

Kort W., Schittekatte M., & Compaan E. (2008). *Clinical Evaluation of Language Fundamentals*, Fourth Edition in Dutch [Measurement instrument]. Amsterdam, The Netherlands: Pearson.

Koster, C., Been, P.H., Krikhaar, E.M., Zwarts, F., Diepstra, H.D., & Van Leeuwen, T.H. (2005). Differences at 17 months: Productive language patterns in infants at familial risk for dyslexia and typically developing infants. *Journal of Speech, Language and Hearing Research, 48*(2), 426–438.

Laasonen, M., Väre, J., Oksanen-Hennah, H., Leppämäki, S., Tani, P., Harno, H., Hokkanen, L., Pothos, E., Cleeremans, A. (2014). Project DyAdd: Implicit learning in adult dyslexia and ADHD. *Annals of Dyslexia, 64*(1), 1–33.

Lammertink, I., Boersma, P., Wijnen, F.N.K., & Rispens, J. (2017). Statistical learning in specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 60*(12), 3474–3486.

Lammertink, I., Boersma, P., Wijnen, F.N.K., & Rispens, J.E. (2019a). Children with developmental language disorder have an auditory verbal statistical learning deficit: Evidence from an online measure. *Language Learning*, Advance online publication: DOI: 10.1111/lang.12373

Lammertink, I., Boersma, P., Wijnen, F.N.K., & Rispens, J.E. (2019b). 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. *Poster presented at the Interdisciplinary Approaches to Statistical Learning (IASL)*, San Sebastian, Spain. DOI: 10.13140/RG.2.2.23251.53285

Lammertink, I.L., Boersma, P.P.G., Wijnen, F.N.K., & Rispens, J.E. (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.*

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

Law, J.M., Vandermosten, M., Ghesquière, P., & Wouters, J. (2017). Predicting future reading problems based on pre-reading auditory measures: A

longitudinal study of children with a familial risk of dyslexia. *Frontiers in Psychology, 8*, Article 124.

Le Clercq, C.M., van der Schroeff, M.P., Rispens, J.E., Ruytjens, L., Goedegebure, A., van Ingen, G., & Franken, M.C. (2017). Shortened nonword repetition task (NWR-S): A simple, quick, and less expensive outcome to identify children with combined specific language and reading impairment. *Journal of Speech, Language, and Hearing Research, 60*(8), 2241–2248.

Leonard, L.B. (2014). *Children with specific language impairment*. Cambridge, United States: MIT press.

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.

Lum, J.A., Conti-Ramsden, G., Morgan, A.T., & Ullman, M.T. (2014). Procedural learning deficits in specific language impairment (SLI): A meta-analysis of serial reaction time task performance. *Cortex, 51*, 1–10.

Lum, J.A., Conti-Ramsden, G., Page, D., & Ullman, M.T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex, 48*(9), 1138–1154.

Lum, J.A., & 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.

Lum, J.A., Kidd, E., Davis, S., Conti-Ramsden, G. (2010). Longitudinal study of declarative and procedural memory in primary school-aged children. *Australian Journal of Psychology, 62*(3), 139–148.

Lum, J.A., Ullman, M.T., & 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*(10), 3460–3476.

Lyytinen, H., Ahonen, T., Eklund, K., Guttorm, T.K., Laakso, M.L., Leinonen, S., Leppanen, P.H.T., Lyytinen, P., Poikkeus, A.-M., Puolakanaho, A., Richardson, U., & Viholainen, H. (2001). Developmental pathways of children with and without familial risk for dyslexia during the first years of life. *Developmental Neuropsychology, 20*(2), 535–554.

Mann, V.A., Shankweiler, D., & Smith, S.T. (1984). The association between comprehension of spoken sentences and early reading ability: The role of phonetic representation. *Journal of Child Language, 11*(3), 627–643.

Marcus, G.F., Pinker, S., Ullman, M.T., Hollander, M., Rosen, T.J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development,* 57(4), 1–182.

Mavridis, D., & Salanti, G. (2014). Exploring and accounting for publication bias in mental health: A brief overview of methods. *Evidence-Based Mental Health, 17*(1), 11-15.

Maybery, M., Taylor, M., & O'Brien-Malone, A. (1995). Implicit learning: Sensitive to age but not IQ. *Australian Journal of Psychology, 47*(1), 8–17.

McArthur, G.M., Hogben, J.H., Edwards, V.T., Heath, S.M., & Mengler, E.D. (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.

McGregor, K.K. (2017). Semantics in child language disorders. In R.G. Schwartz (ed.), *The handbook of child language disorders* (pp. 392–415). New York, United States: Routledge.

Melby-Lervåg, M., Lyster, S.A., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin, 138*(2), 322–352.

Menghini, D., Hagberg, G.E., Caltagirone, C., Petrosini, L., & Vicari, S. (2006). Implicit learning deficits in dyslexic adults: An fMRI study. *NeuroImage, 33*(4), 1218–1226.

Menghini, D., Finzi, A., Benassi, M., Bolzani, R., Facoetti, A., Giovagnoli, S., Ruffino, M., Vicari, S. (2010). Different underlying neurocognitive deficits in developmental dyslexia: A comparative study. *Neuropsychologia, 48*(4), 863–872.

Miles, T.R. (2004). Some problems in determining the prevalence of dyslexia. *Electronic Journal of Research in Educational Psychology, 2*(2), 5–12.

Misyak, J.B., & Christiansen, M.H. (2012). Statistical learning and language: An individual differences study. *Language Learning, 62*(1), 302–331.

Misyak, J.B., Christiansen, M.H., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology, 1*(31), 1–9.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G. (2009). Academia and clinic annals of internal medicine preferred reporting items for systematic reviews and meta-analyses. *Annuals of Internal Medicine, 151*(4), 264–269.

Nash, H.M., Hulme, C., Gooch, D., & Snowling, M.J. (2013). Preschool language profiles of children at family risk of dyslexia: Continuities with specific language impairment. *Journal of Child Psychology and Psychiatry, 54*(9), 958–968.

Nicolson, R.I., & Fawcett, A.J. (2007). Procedural learning difficulties: Reuniting the developmental disorders? *Trends in Neurosciences, 30*(4), 135–141.

Nicolson, R.I., & Fawcett, A.J. (2011). Dyslexia, dysgraphia, procedural learning and the cerebellum. *Cortex, 47*(1), 117–127.

Nigro, L., Jiménez-Fernández, G., Simpson, I.C., & Defior, S. (2015). Implicit learning of written regularities and its relation to literacy acquisition in a shallow orthography. *Journal of Psycholinguistic Research, 44*(5), 571–585.

Nigro, L., Jiménez-Fernández, G., Simpson, I.C., & Defior, S. (2016). Implicit learning of non-linguistic and linguistic regularities in children with dyslexia. *Annals of Dyslexia, 66*(2), 1–17.

Nissen, M.J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*(1), 1–32.

Norton, E.S., & Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annual Review of Psychology, 63*, 427–452.

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

Orgassa, A., & Weerman, F. (2008). Dutch gender in specific language impairment and second language acquisition. *Second Language Research, 24*(3), 333–364.

Pacton, S., Fayol, M., & Perruchet, P. (2005). Children's implicit learning of graphotactic and morphological regularities. *Child Development, 76*(2), 324–339.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit Learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General, 130*(3), 401–426.

Papadopoulos, T.C., Spanoudis, G.C., & Georgiou, G.K. (2016). How is RAN related to reading fluency? A comprehensive examination of the prominent theoretical accounts. *Frontiers in Psychology, 7*, Article 1217.

Pavlidou, E.V., Kelly, L.M., & Williams, J.M. (2010). Do children with developmental dyslexia have impairments in implicit learning? *Dyslexia, 16*(2), 143–161.

Pavlidou, E.V., & Williams, J.M. (2014). Implicit learning and reading: Insights from typical children and children with developmental dyslexia using the

artificial grammar learning (AGL) paradigm. *Research in Developmental Disabilities, 35*(7), 1457–1472.

Pavlidou, E.V., Williams, J.M., & Kelly, L.M. (2009). Artificial grammar learning in primary school children with and without developmental dyslexia. *Annals of Dyslexia, 59*(1), 55–77.

Pennington, B.F. (2006). From single to multiple deficit models of developmental disorders. *Cognition, 101*(2), 385–413.

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

Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine, 26*(25), 4544–4562.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York, United States: Harper Collins.

Plante, E., & Gómez, R.L. (2018). Learning without trying: The clinical relevance of statistical learning. *Language, Speech, and Hearing Services in Schools, 49*(3S), 710–722.

Pothos, E.M., & Kirk, J. (2004). Investigating learning deficits associated with dyslexia. *Dyslexia, 10*(1), 61–76.

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from http://www.pstnet.com

Qi, Z., Sanchez Araujo, Y., Georgan, W.C., Gabrieli, J.D., & Arciuli, J. (2018). Hearing matters more than seeing: A cross-modality study of statistical learning and reading ability. *Scientific Studies of Reading, 23*(1), 101–115.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Ramus, F. (2003). Developmental dyslexia: Specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology, 13*(2), 212–218.

Ramus, F., Pidgeon, E., & Frith, U. (2003). The relationship between motor control and phonology in dyslexic children. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 44*(5), 712–722.

Ramus, F., Marshall, C.R., Rosen, S., & van der Lely, H.K. (2013). Phonological deficits in specific language impairment and developmental dyslexia: Towards a multidimensional model. *Brain, 136*(2), 630–645.

Ramus, F., Rosen, S., Dakin, S.C., Day, B.L., Castellote, J.M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain, 126*(4), 841–865.

Ramus, F., & Szenkovits, G. (2008). What phonological deficit? *The Quarterly Journal of Experimental Psychology, 61*(1), 129–141.

Raven, J. & Raven, J. (2003). Raven progressive matrices [Measurement instrument]. In R. S. McCallum (Ed.) *Handbook of nonverbal assessment* (pp. 223–237). New York, United States: Kluwer Academic/Plenum Publishers.

Raviv, L., & Arnon, I. (2017). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science, 21*(4), Article e12593.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*(6), 855–863.

Reggiani, D. (2010). *Dyslexia and the acquisition of syntax: Passive and control* (Doctoral dissertation). University of Verona, Verona, Italy.

Rice, M.L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech, Language, and Hearing Research, 39*(6), 1239–1257.

Rispens, J.E., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research, 55*(3), 683–694.

Rispens, J.E., Baker, A., & Duinmeijer, I. (2015). Word recognition and nonword repetition in children with language disorders: The effects of neighborhood density, lexical frequency, and phonotactic probability. *Journal of Speech, Language, and Hearing Research, 58*(1), 78–92.

Rispens, J.E., & Been, P. (2007). Subject-verb agreement and phonological processing in developmental dyslexia and specific language impairment (SLI): A closer look. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists, 42*(3), 293–305.

Rispens, J.E., de Bree, E.H., & Kerkhoff, A. (2014). What's in a suffix? The past tense in Dutch children with reading problems. In J. Hoeksema, D. Gilbers

& P. Hendriks (Eds.), *Black book: A festschrift in honor of Frans Zwarts* (pp. 271–281), University of Groningen, The Netherlands.

Rispens, J.E., Roeleven, S., & Koster, C. (2004). Sensitivity to subject–verb agreement in spoken language in children with developmental dyslexia. *Journal of Neurolinguistics, 17*(5), 333–347.

Robertson, E.K., & Joanisse, M.F. (2010). Spoken sentence comprehension in children with dyslexia and language impairment: The roles of syntax and working memory. *Applied Psycholinguistics, 31*(1), 141–165.

Robertson, E.M. (2007). The serial reaction time task: Implicit motor skill learning? *Journal of Neuroscience, 27*(38), 10073–10075.

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

Roodenrys, S., & Dunn, N. (2008). Unimpaired implicit learning in children with developmental dyslexia. *Dyslexia, 14*(1), 1–15.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Rüsseler, J., Gerth, I., & Münte, T.F. (2006). Implicit learning is intact in adult developmental dyslexic readers: Evidence from the serial reaction time task and artificial grammar learning. *Journal of Clinical and Experimental Neuropsychology, 28*(5), 808–827.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

Samara, A. (2013). *Statistical learning of orthographic patterns in typically developing and dyslexic populations* (Doctoral dissertation). Bangor University, Bangor, United Kingdom.

Samara, A., Singh, D., & Wonnacott, E. (2019). Statistical learning and spelling: Evidence from an incidental learning experiment with children. *Cognition, 182*, 25–30.

Scarborough, H.S. (1990). Very early language deficits in dyslexic children. *Child Development, 61*(6), 1728–1743.

Schiff, R., & Katan, P. (2014). Does complexity matter? Meta-analysis of learner performance in artificial grammar tasks. *Frontiers in Psychology, 5*, Article 1048.

Schittekatte, M., Groenvynck, H., Fontaine, J., & Dekker, P. (2007). *TEAch: Test of Everyday Attention for Children*, Dutch version [Measurement instrument]. Amsterdam, The Netherlands: Pearson.

Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia, 67*(2), 147-162.

Schmalz, X., Moll, K., Mulatti, C., & Schulte-Körne, G. (2019). Is statistical learning ability related to reading ability, and if so, why? *Scientific Studies of Reading, 23*(1), 64–76.

Schneider, W., Eschman, A., and Zuccolotto, A. (2012). *E-Prime user's guide*. Psychology Software Tools, Inc.

Schwarzer, G., Carpenter, J., & Rücker, G. (2010). Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis. *Journal of Clinical Epidemiology, 63*(3), 282–288.

Schwarzer, G. (2012). *Package "meta": General package for meta-analysis* [R package version 4.9-7], http://CRAN.R-project.org/package=meta

Shankweiler, D., Crain, S., Katz, L., Fowler, A.E., Liberman, A.M., Brady, S.A., Thornton, E., Lundquist, E., Dreyer, L., Fletcher, J.M., Stuebing, K.K., Shaywitz, S.E., & Shaywitz, B.A. (1995). Cognitive profiles of reading-disabled children: Comparison of language skills in phonology, morphology, and syntax. *Psychological Science, 6*(3), 149–156.

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science, 42*(8), 3100–3115.

Siegel, L.S. (2006). Perspectives on dyslexia. *Paediatrics and Child Health, 11*(9), 581–587.

Siegelman, N., Bogaerts, L., Armstrong, B.C., & Frost, R. (2019). What exactly is learned in visual statistical learning? Insights from Bayesian modeling. *Cognition, 192*, Article 104002.

Siegelman, N., Bogaerts, L., & Frost, R. (2017a). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49*(2), 418–432.

Siegelman N., Bogaerts L., Christiansen M.H., & Frost R. (2017b). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), Article 20160059.

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*(S3), 692–727.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120.

Sigurdardottir, H.M., Danielsdottir, H.B., Gudmundsdottir, M., Hjartarson, K.H., Thorarinsdottir, E.A., & Kristjánsson, Á. (2017). Problems with visual statistical learning in developmental dyslexia. *Scientific Reports, 7*(1), Article 606.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Singh, S., Walk, A.M., & Conway, C.M. (2018). Atypical predictive processing during visual statistical learning in children with developmental dyslexia: An event-related potential study. *Annals of Dyslexia, 68*(2), 1–15.

Snowling, M.J. (2000) *Dyslexia*. Oxford, England: Blackwell.

Snowling, M.J. (2001). From language to reading and dyslexia. *Dyslexia, 7*(1), 37–46.

Snowling, M.J., & Melby-Lervåg, M. (2016). Oral language deficits in familial dyslexia: A meta-analysis and review. *Psychological Bulletin, 142*(5), 498–545.

Spencer, M., Kaschak, M.P., Jones, J.L., & Lonigan, C.J. (2014). Statistical learning is related to early literacy-related skills. *Reading and Writing, 28*(4), 467–490.

Sperling, A.J., Lu, Z.L., & Manis, F.R. (2004). Slower implicit categorical learning in adult poor readers. *Annals of Dyslexia, 54*(2), 281–303.

Staels, E., & Van den Broeck, W. (2017). A specific implicit sequence learning deficit as an underlying cause of dyslexia? Investigating the role of attention in implicit learning tasks. *Neuropsychology, 31*(4), 371–382.

Steacy, L.M., Compton, D.L., Petscher, Y., Elliott, J.D., Smith, K., Rueckl, J.G., Sawi, O., Frost, S.J., & Pugh, K.R. (2018). Development and prediction of context-dependent vowel pronunciation in elementary readers. *Scientific Studies of Reading, 23*(1), 49–63.

Stein, C.L., Cairns, H.S., & Zurif, E.B. (1984). Sentence comprehension limitations related to syntactic deficits in reading-disabled children. *Applied Psycholinguistics, 5*(04), 305–322.

Stein, J.F., & Walsh, V. (1997). To see but not to read; The magnocellular theory of dyslexia. *Trends in Neurosciences, 20*(4), 147–152.

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

Swanson, H.L., & Jerman, O. (2007). The influence of working memory on reading growth in subgroups of children with reading disabilities. *Journal of Experimental Child Psychology, 96*(4), 249–283.

Szmalec, A., Loncke, M., Page, M.P. a, & Duyck, W. (2011). Order or disorder? Impaired Hebb learning in dyslexia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1270–1279.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. New York, United States: Pearson.

Tallal, P. (2004). Improving language and literacy is a matter of time. *Nature Reviews Neuroscience, 5*(9), 721–728.

Thompson, S.G., & Higgins, J. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*(11), 1559–1573.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, United States: Harvard University Press.

Toro, J.M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*(2), B25–B34.

Treiman, R. (2018). Statistical learning and spelling. *Language, Speech, and Hearing Services in Schools, 49*(3S), 644–652.

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

Ullman, M.T. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition, 4*(2), 105–122.

Ullman, M.T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*(1-2), 231–270.

Ullman, M.T., & Pierpont, E.I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex, 41*(3), 399–433.

Ullman, M.T., Earle, F.S., Walenski, M., & Janacsek, K. (2019). The neurocognition of developmental disorders of language. *Annual Review of Psychology, 71*(5), 389–417.

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.

van Alphen, P., de Bree, E.H., Gerrits, E., De Jong, J., Wilsenach, C., & Wijnen, F.N.K. (2004). Early language development in children with a genetic risk of dyslexia. *Dyslexia, 10*(4), 265–288.

van den Bos, K.P., & Spelberg, H. (2007) *Continu benoemen & woorden lezen*. Amsterdam, The Netherlands: Boom uitgevers Amsterdam.

van den Bos, K.P., Spelberg, H., Scheepsma, A., & de Vries, J. (1994). *De Klepel. Vorm A en B. Een test voor de leesvaardigheid van pseudowoorden. Verantwoording, handleiding, diagnostiek en behandeling*. Nijmegen, The Netherlands: Berkhout.

van Elk, M., Matzke, D., Gronau, Q., Guang, M., Vandekeckhove, J., & Wagenmakers, E.J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology, 6*, Article 1365.

van der Kleij, S.W., Groen, M.A., Segers, E., & Verhoeven, L. (2019). Sequential implicit learning ability predicts growth in reading skills in typical readers and children with dyslexia. *Scientific Studies of Reading, 23*(1), 77–88.

van Setten, E.R., Tops, W., Hakvoort, B.E., van der Leij, A., Maurits, N.M., & Maassen, B.A. (2017). L1 and L2 reading skills in Dutch adolescents with a familial risk of dyslexia. *PeerJ: Brain and Cognition*, *5*, Article e3895.

van Witteloostuijn, M.T.G., Boersma, P., Wijnen, F.N.K., & Rispens, J.E. (2017). Visual artificial grammar learning in dyslexia: A meta-analysis. *Research in Developmental Disabilities, 70*, 126–137.

van Witteloostuijn, M.T.G., Boersma, P., Wijnen, F.N.K., & Rispens, J.E. (2019). Statistical learning abilities of children with dyslexia across three experimental paradigms. *PLoS ONE, 14*(8), Article e0220041.

van Witteloostuijn, M.T.G., Boersma, P., Wijnen, F.N.K., & Rispens, J.E. (under review). The contribution of individual differences in statistical learning to reading and spelling performance in children with and without dyslexia. *Dyslexia*.

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

van Witteloostuijn, M.T.G., Lammertink, I., Boersma, P., Wijnen, F.N.K., & Rispens, J.E. (2019). Assessing visual statistical learning in early-school-aged children: The usefulness of an online reaction time measure. *Frontiers in Psychology, 10*, Article 2051.

Vellutino, F.R., Fletcher, J.M., Snowling, M.J., & Scanlon, D.M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*(1), 2–40.

Verhoeven, L., Steenge, J., & van Balkom, H. (2011). Verb morphology as clinical marker of specific language impairment: Evidence from first and second language learners. *Research in Developmental Disabilities, 32*(3), 1186–1193.

Vicari, S., Marotta, L., Menghini, D., Molinari, M., & Petrosini, L. (2003) Implicit learning deficit in children with developmental dyslexia. *Neuropsychologia, 41*(1), 108–114.

Vicari, S., Finzi, A., Menghini, D., Marotta, L., Baldi, S., & Petrosini, L. (2005). Do children with developmental dyslexia have an implicit learning deficit? *Journal of Neurology, Neurosurgery, and Psychiatry, 76*(10), 1392–1397.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Von Koss Torkildsen, J., Arciuli, J., & Wie, O.B. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences, 69*, 60–68.

Waber, D.P., Marcus, D.J., Forbes, P.W., Bellinger, D.C., Weiler, M.D., Sorensen, L.G., & Curran, T. (2003). Motor sequence learning and reading ability: Is poor reading associated with sequencing deficits? *Journal of Experimental Child Psychology, 84*(4), 338–354.

Wagenmakers, E.J., Wetzels, R., Borsboom, D., van der Maas, H.L., & Kievit, R.A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638.

Waltzman, D., & Cairns, H. (2000). Grammatical knowledge of third grade good and poor readers. *Applied Psycholinguistics, 21*(2), 263–284.

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

West, G., Shanks, D., & Hulme, C. (2018). Sustained attention, not procedural learning, is a predictor of language, reading and arithmetic skills in children. Preprint DOI: http://dx.doi.org/10.31234/osf.io/afrms

West, G., Clayton, F.J., Shanks, D.R., & Hulme, C. (2019). Procedural and declarative learning in dyslexia. *Dyslexia, 25*(3), 246–255.

Wexler, K. (1994). Optional infinitives, head movement, and the economy of derivation. In D. Lightfoot, & N. Hornstein (Eds.), *Verb movement*. New York, United States: Cambridge University Press.

Wijnen, F.N.K. (2013). Acquisition of linguistic categories: Cross-domain convergence. In J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech, and language: Exploring the evolution of mind and brain*. Cambridge, United States: MIT Press.

Willingham, D.B., Nissen, M.J., and Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1047–1060.

Wolf, M., & Bowers, P.G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91*(3), 415–438.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development, 4*(1), 32–62.

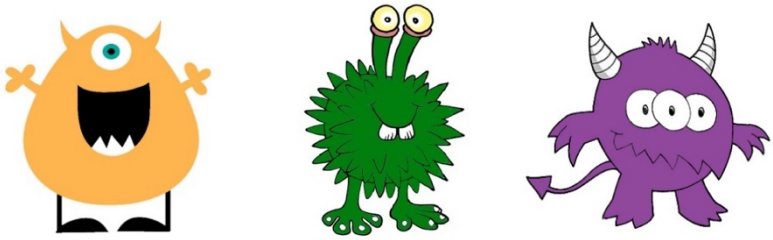
Appendices

Overview

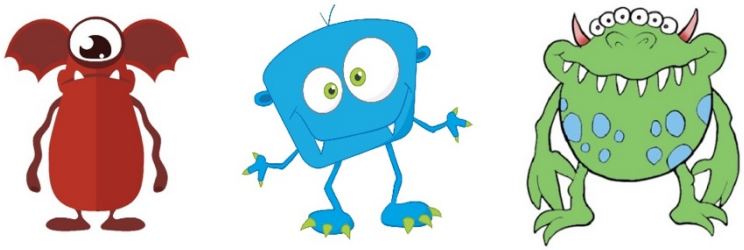
	Chapter	Contents
Appendix A	Chapter 2	VSL stimuli
Appendix B	Chapter 2	VSL test items
Appendix C	Chapter 2	VSL instructions
Appendix D	Chapter 2	Supplementary figure
Appendix E	Chapter 3	Search terms for databases
Appendix F	Chapter 3	Extracted data per study
Appendix G	Chapter 4	A-NADL *X*- and *f*-elements
Appendix H	Chapter 4	A-NADL on- and offline task

Appendix A – Chapter 2, VSL stimuli

Triplet ABC

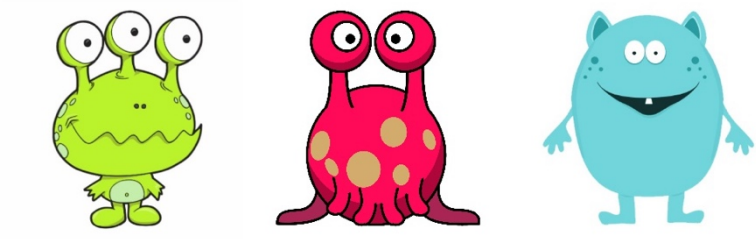


Triplet DEF



Triplet GHI

Triplet JKL



Appendix B – Chapter 2, VSL test items – I

2-AFC: pattern recognition

Item	Grammatical	Ungrammatical	Chance	Similarity distractor
1	**ABC**	DHL	50%	No
2	**ABC**	GKC	50%	Yes
3	**DEF**	GKC	50%	No
4	**DEF**	JBF	50%	Yes
5	**GHI**	JBF	50%	No
6	**GHI**	AEI	50%	Yes
7	**JKL**	AEI	50%	No
8	**JKL**	DHL	50%	Yes
9	**AB**	JB	50%	Yes
10	**AB**	DH	50%	No
11	**BC**	BF	50%	Yes
12	**BC**	EI	50%	No
13	**DE**	AE	50%	Yes
14	**DE**	KC	50%	No
15	**EF**	BF	50%	Yes
16	**EF**	GK	50%	No
17	**GH**	DH	50%	Yes
18	**GH**	EI	50%	No
19	**HI**	HL	50%	Yes
20	**HI**	GK	50%	No
21	**JK**	JB	50%	Yes
22	**JK**	HL	50%	No
23	**KL**	KC	50%	Yes
24	**KL**	AE	50%	No

Appendix B – Chapter 2, VSL test items – II

3-AFC: pattern completion				
Item	**Pattern**	**Answer options**	**Chance**	**Position question mark**
25	A ? C	**B**, J, H	33%	2
26	D E ?	**F**, B, C	33%	3
27	? H I	**G**, L, A	33%	1
28	J ? L	**K**, F, E	33%	2
29	? B C	**A**, H, G	33%	1
30	D ? F	**E**, G, B	33%	2
31	G H ?	**I**, D, L	33%	3
32	? K L	**J**, C, E	33%	1
33	B ?	**C**, E, F	33%	1
34	? C	**B**, D, K	33%	2
35	? E	**D**, K, J	33%	1
36	E ?	**F**, G, C	33%	3
37	G ?	**H**, A, K	33%	2
38	H ?	**I**, D, L	33%	3
39	? K	**J**, I, A	33%	1
40	K ?	**L**, H, I	33%	3

Appendix C – Chapter 2, VSL instructions

General instructions

Dutch: *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.*

English: 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. Afterwards, you will automatically see the next alien standing in the line.

Dutch: *In dit spel vinden sommige aliens elkaar heel leuk. Zij staan bij elkaar in de rij. Bekijk elke alien goed en let goed op de volgorde van de aliens, want daarover stel ik je later nog wat vragen.*

English: 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.

Cover task instructions

Dutch: *Dit is een indringer! De indringer mag niet mee op het ruimteschip. Als je deze indringer ziet, moet je hem wegjagen. Dit doe je door op hem te drukken. Je kan gewoon met je vinger op het scherm drukken. Probeer maar!*

English: This is an intruder! The intruder is not allowed to join the others on the spaceship. If you see this intruder, you have to scare him away. You can do this by touching him on the screen with your finger. Try it!

Dutch: *Goed zo! Als je de indringer weggejaagd hebt, gaat het spel verder.*

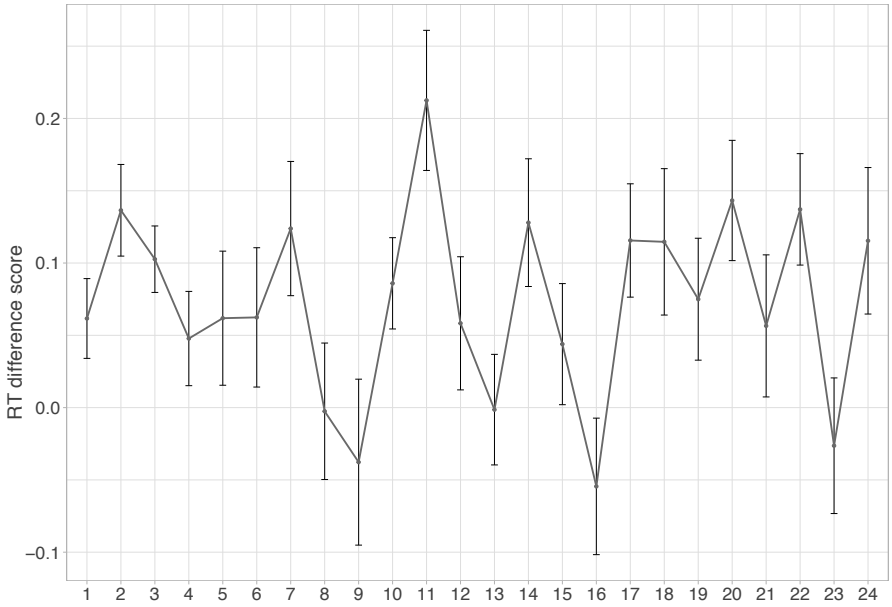
English: Well done! When you scare away the intruder, the game continues.

Test phase instructions

Dutch: *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.*

English: Now we're going to do 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 this.

Appendix D – Chapter 2, VSL supplementary figure



Supplementary figure: Descriptive results of the online RT data: difference score. Mean normalized RT to unpredictable element 1 minus mean normalized RT to predictable element 2, plotted per repetition of triplets during the experiment (see §3.1.2 of the manuscript).

Appendix E – Chapter 3, Search terms for databases

Search terms for all databases consisted of two parts associated with the two most important selection criteria: a) procedural learning, and b) participants with dyslexia. Search terms are presented for each individual database below.

1. Ovid (used to search PsycInfo, Medline, and ERIC)
    a. ("procedural learning" or "implicit learning" or "artificial grammar learning" or "statistical learning"), ti, ab.
    b. (dyslexia or "reading difficulties" or "reading disorder" or "reading disability" or "spelling difficulties" or "spelling disorder" or "spelling disability"), ti, ab.
    c. a and b.
2. CINAHL: AB ("procedural learning" or "implicit learning" or "artificial grammar learning" or "statistical learning") AND AB (dyslexia or "reading difficulties" or "reading disorder" or "reading disability" or "spelling difficulties" or "spelling disorder" or "spelling disability").
3. LLBA: ab("procedural learning" or "implicit learning" or "artificial grammar learning" or "statistical learning") AND ab(dyslexia or "reading difficulties" or "reading disorder" or "reading disability" or "spelling difficulties" or "spelling disorder" or "spelling disability").
4. Pubmed: (("procedural learning"[Title/Abstract] OR "implicit learning"[Title/Abstract] OR "artificial grammar learning"[Title/Abstract] OR "statistical learning"[Title/Abstract]) AND (dyslexia[Title/Abstract] OR "reading difficulties"[Title/Abstract] OR "reading disorder"[Title/Abstract] OR "reading disability"[Title/Abstract] OR "spelling difficulties"[Title/Abstract] OR "spelling disorder"[Title/Abstract] OR "spelling disability"[Title/Abstract])).
5. OATD: abstract:(("procedural learning" OR "implicit learning" OR "artificial grammar learning" OR "statistical learning") AND (dyslexia OR "reading difficulties" OR "reading disorder" OR "reading disability" OR "spelling difficulties" OR "spelling disorder" OR "spelling disability").

Appendix F – Chapter 3, Extracted data per study – I

Study	Data abstracted for effect size calculation
Du (2013)	Means and standard deviations of the percentage of grammatical and non-grammatical items endorsed were extracted from Table 7 (p. 116). These were later converted to a mean and *SD* of the overall accuracy as described under section 2.4.
Ise et al. (2012)	Means and standard deviations of the percentage of correct classifications of test items in the consonant-vowel condition (experiment a in the present meta-analysis, p. 1005) and consonant-only condition (experiment b in the present meta-analysis, p. 1005) were reported in the text.
Kahta & Schiff **(**2016)	We extracted means and 95% confidence intervals of the percentage of endorsements of grammatical and non-grammatical test items from Figure 2 (p. 241). This was done using DigitizeIt digitizer software. These values were later converted to a mean and *SD* of the overall accuracy as described under section 2.4.
Laasonen et al. (2014)	Marja Laasonen provided us with an Excel sheet including the means and standard deviations of the overall accuracy (in percentage) on the test phase.
Nigro et al. (2016)	Means and standard deviations of the percentage correct on the unseen[a] items in the test phase were reported in text both for the experiment using abstract shapes (experiment a in the present meta-analysis, p. 208-209) and the experiment using letters (experiment b in the present meta-analysis, p. 211).

Note. [a] Performance on unseen items, instead of overall performance on seen and unseen items, was selected for inclusion in this study, as other studies included only novel, unseen items in the test phase. This table continues on the following page.

Appendix F – Chapter 3, Extracted data per study – II

Study	Data abstracted for effect size calculation
Pavlidou et al. (2009; 2010; 2014)	The means and standard deviations of the number of correct responses in the test phase were provided in the text (2009: p. 63, 2010: p. 3292, 2014: p. 1465). These were later converted to the means and standard deviations of the percentage of correct responses.
Pothos & Kirk (2004)	Means and standard deviations of the average performance on the "sequences" test phase were extracted from Figure 4 (p. 71) using DigitizeIt digitizer software.
Rüsseler et al. (2006)	Table 2 (p. 820) presents the percentage of correct classification of test items per individual participant. Means and standard deviations were calculated based on these raw data.
Samara (2013)	Means and standard deviations of the percentage of low chunk strength (non-grammatical) and high chunk strength (grammatical) items endorsed by participants were extracted from Table 4.2 (p. 143). These were later converted to a mean and *SD* of the overall accuracy as described under section 2.4.

Appendix G – Chapter 4, A-NADL *X*- and *f*-elements

***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

Appendix H – Chapter 4, A-NADL on- and offline task

Rule blocks

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

Summary

Examining the contribution of statistical learning to grammar and literacy acquisition: A study of Dutch children with and without dyslexia

Children typically acquire their mother tongue relatively quickly and seemingly effortlessly. Although their language and literacy skills will continue to develop into adulthood, children already know approximately 1,500 words, and are able to combine words into sentences, before they begin primary school at the age of four. So how do they do it? This dissertation investigated the hypothesis that a domain-general learning mechanism supports the acquisition of language, both in its spoken and in its written form. Such a domain-general learning mechanism allows for the learning of abstract patterns and rules based on the statistical properties of the input (i.e. language). This ability to learn from statistical patterns in the world around us is known as "statistical learning", and is thought to be an implicit learning mechanism (e.g. Frost, Armstrong, Siegelman, & Christiansen, 2015; Perruchet & Pacton, 2006; Saffran, Newport, & Aslin, 1996). Moreover, statistical learning has been hypothesized to play an important role in the acquisition of language and literacy skills. Evidence in support of this hypothesized relationship comes from studies that have reported significant correlations between performance on tasks that assess statistical learning ability on the one hand, and measures of performance on language and literacy skills on the other hand (e.g. Arciuli & Simpson, 2012; Frost, Siegelman, Narkiss, & Afek, 2013; Kidd, 2012; Kidd & Arciuli, 2016; Misyak, Christiansen, & Tomblin, 2010).

For some children, however, the acquisition of language and literacy skills is not quick and effortless. Between 3 to 10 percent of the general population is diagnosed with developmental dyslexia (henceforth "dyslexia"; Miles, 2004; Siegel, 2006). Although dyslexia is most commonly characterized by difficulties with (technical) reading and spelling, more subtle problems with spoken language are also observed. These spoken language difficulties may include problems with inflectional morphology (e.g. forming the correct plural

or past tense form; de Bree & Kerkhoff, 2014; Joanisse, Manis, Keating, & Seidenberg, 2000) and syntax (e.g. word order in complex structures such as passives: Reggiani, 2010; Shankweiler et al., 1995). These language and literacy deficits are attributed to dyslexia only if they do not stem from low intelligence, the absence of academic or social opportunities, or sensory or neurological impairments (e.g. DSM-V, 2013). An important aim of dyslexia research is, therefore, to investigate the underlying cause for these deficits. This is where statistical learning comes in: whereas most previous research has focussed on causes for dyslexia in the area of phonology and phonological memory (e.g. de Bree, 2007; Ramus, 2003), more recent studies have suggested that a problem with a domain-general learning mechanism – i.e. statistical learning – may be the underlying cause of dyslexia (e.g. Nicolson & Fawcett, 2007; 2011; Ullman, 2004; Ullman, Sayako, Earle, Walenski, & Janacsek, 2019). If, as previously hypothesized, statistical learning plays a crucial role in the acquisition of spoken and written language in typical development, a deficit in statistical learning may result in problems with language and literacy acquisition, as is the case for individuals with dyslexia. In line with the view that dyslexia stems from a statistical learning deficit, studies have shown that individuals with dyslexia perform worse on statistical learning tasks than their typically developing (TD) peers (e.g. Gabay, Thiessen, & Holt, 2015; Jiménez-Fernández, Vaquero, Jiménez, & Defior, 2011; Lum, Ullman, & Conti-Ramsden, 2013; Pavlidou & Williams, 2014).

Thus, the hypothesized relationship between statistical learning and (language and) literacy skills has been investigated through two separable lines of research: (1) the study of the correlation between individual differences in statistical learning ability and (language and) literacy scores, and (2) the study of group differences between individuals with and without dyslexia. It is important to note here that there have also been reports of null results, both in relation to individual differences (e.g. Schmalz, Moll, Mulatti, & Schulte-Körne, 2019; West, Vadillo, Shanks, & Hulme, 2017), and with repect to group differences (e.g. Kelly, Griffiths, & Frith, 2002; Kerkhoff, de Bree, de Klerk, & Wijnen, 2013). This mixed pattern of findings in the field (i.e. some studies reporting significant results, some reporting null results) has led to questions regarding the strength of the relationship between statistical learning and spoken and written language, but also to questions relating to our ability to assess this relationship reliably,

especially in children (e.g. Arnon, 2019b; Kidd, Donnelly, & Christiansen, 2017; Schmalz et al., 2019; West et al., 2017). There is a need for large-scale studies that comprehensively and reliably investigate the relationship between statistical learning, language and literacy skills, and developmental impairments such as dyslexia (e.g. Arciuli & Conway, 2018).

The current dissertation aimed to fill this gap and provide valuable evidence about the relationship between statistical learning and the acquisition of language and literacy skills. We examined the contribution of statistical learning to grammar and literacy skills in Dutch-speaking children with and without a diagnosis of dyslexia. Importantly, the studies presented in this dissertation control for individual variation in cognitive constructs related to statistical learning and/or the acquisition of spoken and written language (e.g. attention, memory, phonological processing; see e.g. Arciuli, 2017). Furthermore, the statistical learning measures used in the present dissertation were developed to more reliably assess the statistical learning abilities of school-aged children. In the next sections, we elaborate on the methodology used and the results obtained in chapters 2 through 6, followed by this dissertation's conclusions and implications for future research.

Prior to commencing our experimental studies of the relationship between statistical learning and spoken and written language skills, we aimed to improve previous statistical learning tasks for use with child participants. In chapter 2, we developed a novel, self-paced visual statistical learning (VSL) task, which was administered among 53 children between 5 and 8 years of age (van Witteloostuijn, Lammertink, Boersma, Wijnen, & Rispens, 2019). Traditionally, learning in such tasks is measured through a two-alternative forced choice (2-AFC) task that is administered after exposure to the VSL structure. This type of measures, that take place after the exposure phase, is referred to as "offline" measures. The VSL task introduced in chapter 2 is based on a previous study with adults and included a novel reaction time (RT) measure. Importantly, this RT measure assesses learning during exposure to the VSL structure (i.e. it is an "online" measure, Siegelman, Bogaerts, Christiansen, & Frost, 2018). The development of online measures is essential, since the reliability of 2-AFC questions has been questioned (e.g. Siegelman & Frost, 2015) and offline measures do not inform us about the learning process during exposure (e.g. Siegelman, Bogaerts, & Frost, 2017). Chapter 2 was the first study to demonstrate the usefulness of such an online RT measure in the VSL task among early-school-

aged children, as was previously demonstrated for adults (Siegelman et al., 2018). The subsequent offline test phase showed no evidence of sensitivity to the VSL structure in 5- to 8-year-old children. Thus, chapter 2 highlighted the potential of online measures of statistical learning, also for use in child participants.

Chapters 3 and 4 examined the statistical learning performance of individuals with dyslexia: while chapter 3 presented a meta-analysis of previous studies using the visual artificial grammar learning (AGL) paradigm (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017), chapter 4 adopted novel measures of statistical learning in an experimental study (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019). In our meta-analysis of 13 previous studies that looked at the visual AGL performance of individuals with dyslexia, we found an effect of group on performance: overall, individuals with dyslexia performed worse than TD individuals. However, we also found evidence of a publication bias in the field; it seems likely that statistically significant studies were more likely to be published than statistically non-significant studies. Once we controlled for this publication bias, the effect of group on performance no longer reached significance. Thus, our initial result is likely to be overly optimistic; the reported difference in visual AGL performance between individuals with and without dyslexia may in fact be nulled by unpublished findings.

In chapter 4, we adopted three other statistical learning paradigms to gain further insight into the question whether individuals with dyslexia experience problems in this area. Our sample comprised 50 8- to 11-year-old children with a prior diagnosis of dyslexia and 50 individually age-matched control TD children. Besides using our novel self-paced VSL task (van Witteloostuijn, Lammertink et al., 2019, Chapter 2), we used a novel online measure of learning in a so-called "auditory nonadjacent dependency learning" (A-NADL) task (Lammertink, van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019) and a more traditional visuo-motoric serial reaction time (SRT) task. In doing so, we aimed to test statistical learning across domains (visual, auditory, and visuo-motoric) and across types of statistical structure (e.g. adjacent and nonadjacent [i.e. more distant] structures). Furthermore, we controlled for sustained attention and short-term and working memory, since these cognitive capacities are thought to relate to statistical learning performance (e.g. Arciuli, 2017; Baker, Olson, & Behrmann, 2004) and individuals with dyslexia are known to experience problems in these areas (e.g. Bosse, Tainturier, & Valdois, 2007; Cowan et al.,

2017). Our findings across the three statistical learning tasks were consistent: we found evidence of sensitivity to the statistical structures in all measures when collapsing over children with and without dyslexia. Also consistently, we found no evidence of a difference in performance between children with and without dyslexia on the VSL, A-NADL and/or SRT tasks. Therefore, the results of chapter 4 (and 3) do not support the hypothesisis that a domain-general statistical learning deficit underlies the (language and) literacy problems that we see in individuals with dyslexia.

Although there was no evidence for (or against) a difference in statistical learning performance between children with dyslexia and those without, it could still be the case that individual differences in statistical learning ability contribute to variation in language and literacy performance. This was hypothesized, since (1) children with and without dyslexia were shown to learn in all three tasks, and (2) the relationship between statistical learning and (spoken and written) language also exists in typical populations. In chapters 5 and 6, we inspected the individual differences in statistical learning of our sample of 100 children with and without dyslexia combined, and we associated these individual differences with children's performance on tasks that assessed grammatical skills (inflectional morphology and syntax; van Witteloostuijn, Boersma, Wijnen, & Rispens, submitted, see chapter 6) and literacy attainment (technical reading, spelling; van Witteloostuijn, Boersma, Wijnen, & Rispens, under review, see chapter 5). Importantly, we calculated the internal consistency and reliability of our statistical learning tasks through the split-half reliability measure (see also Arnon, 2019a; Siegelman, Bogaerts, Christiansen, & Frost, 2017b), and controlled for a range of participant-level variables (e.g. age, attention, memory, phonological processing). Again, the results from chapters 5 and 6 are consistent: we find no support for (or against) a relationship between statistical learning on the one hand, and performance on grammar (inflectional morphology and syntax) and literacy (technical reading and spelling) measures on the other hand. The split-half reliabilities of our statistical learning measures were found to approach the standard for psychometric testing of .80 (ranging between .70 on the VSL offline measure and .79 on the A-NADL online measure). Exploratory findings do support a link between phonological skills (phonological processing, phonological short-term and working memory) and both grammar and literacy skills in children with and without dyslexia. Finally, chapter 6 highlighted the subtle problems that children with dyslexia experience with grammar. These

problems surfaced both in the area of inflectional morphology (e.g. irregular plural and past tense formation), and the area of syntax (e.g. verb second word order).

Taken together, the results presented in this dissertation do not provide evidence for (or against) a link between a domain-general statistical learning ability and the acquisition of language and literacy skills. These results come from two lines of reseach: (1) we found no evidence that individual differences in statistical learning contribute to performance on tasks that measure language and literacy attainment in children with and without dyslexia, and (2) although children, overall, were shown to be sensitive to the statistical structures presented to them in our statistical learning tasks, we found no evidence for (or against) a difference in performance between children with and without dyslexia. To conclude, it cannot be excluded that the relationship between statistical learning and language and literacy acquisition may be less strong than hypothesized and may only surface when specific methodological choices are made. Furthermore, individuals with dyslexia likely do not have a domain-general, extensive deficit in statistical learning. More research in the form of large-scale and pre-registered studies, as well as meta-analyses, is needed in order to reach definitive conclusions regarding the contribution of (domain-general) statistical learning ability to the acquisition of language and literacy skills, both in typical and in impaired populations.

Samenvatting

De bijdrage van statistisch leren aan de verwerving van grammatica en lees- en spelvaardigheid: Een studie van Nederlandse kinderen met en zonder dyslexie

Kinderen verwerven gewoonlijk hun moedertaal snel en zonder al te veel moeite. Hoewel taalvermogen en lees- en schrijfvaardigheid zich blijft ontwikkelen tot in de volwassenheid, kennen kinderen al circa 1.500 woorden wanneer ze beginnen aan de basisschool op de leeftijd van vier jaar. Daarnaast zijn ze in staat om deze woorden te combineren tot grammaticale zinnen zonder dat kinderen dit bewust is aangeleerd. Een belangrijke vraag is: hoe doen kinderen dit? Deze dissertatie onderzocht de hypothese dat een domeinoverstijgend leermechanisme de verwerving van taal ondersteunt, zowel in gesproken als in geschreven vorm. Zo'n domeinoverstijgend leermechanisme draagt bij aan het leren van abstracte patronen en regels op basis van de statistische eigenschappen van "input" (e.g. taal). Deze vaardigheid om statistische patronen in de wereld om ons heen te ontdekken wordt "statistisch leren" genoemd (bijvoorbeeld Frost, Armstrong, Siegelman, & Christiansen, 2015; Saffran, Newport, & Aslin, 1996). Aangenomen wordt dat statistisch leren een impliciet leerproces is, dat plaatsvindt zonder bewuste inspanning (Perruchet & Pacton, 2006). Bovendien wordt verondersteld dat statistisch leren een belangrijke rol speelt in de verwerving van gesproken en geschreven taal. Bewijs voor deze hypothese komt voort uit studies die significante correlaties rapporteren tussen (a) de prestaties op taken die het statistisch leervermogen meten en (b) scores op taken op het gebied van gesproken en geschreven taal (Arciuli & Simpson, 2012; Frost, Siegelman, Narkiss, & Afek, 2013; Kidd, 2012; Kidd & Arciuli, 2016; Misyak, Christiansen, & Tomblin, 2010).

Voor sommige kinderen verloopt de verwerving van gesproken en geschreven taal niet snel of moeiteloos. Tussen de 3 en 10 procent van de algemene bevolking heeft een diagnose ontwikkelingsdyslexie (hierna "dyslexie"; Miles, 2004; Siegel, 2006). Hoewel dyslexie vooral gekenmerkt wordt door moeilijkheden met (technisch) lezen en spellen, is gebleken dat mensen met dyslexie ook subtiele problemen met gesproken taal hebben. Deze problemen

bevinden zich op het gebied van inflectionele morfologie (bijvoorbeeld het vormen van de correcte meervoudsvorm of verledentijdsvorm; de Bree & Kerkhoff, 2014; Joanisse, Manis, Keating, & Seidenberg, 2000) en syntaxis (bijvoorbeeld woordvolgorde in complexe zinnen zoals passieve zinnen; Reggiani, 2010; Shankweiler et al., 1995). Deze problemen met gesproken en geschreven taal bij mensen met dyslexie komen niet voort uit academische (intellectuele) beperkingen, sociale beperkingen, of zintuiglijke of neurologische afwijkingen (DSM-V, 2013). Het onderzoeken van de onderliggende oorzaken voor het ontstaan van deze problemen is daarom een belangrijk doel van wetenschappers op het gebied van dyslexie. Dit is waar statistisch leren in het spel komt: terwijl een groot deel van eerder onderzoek zich heeft gericht op moeilijkheden op het gebied van fonologie en fonologisch geheugen als verklaring voor dyslexie (bijvoorbeeld de Bree, 2007; Ramus, 2003), hebben recente studies gesuggereerd dat een stoornis in een domeinoverstijgend leermechanisme – oftewel statistisch leren – de onderliggende oorzaak van dyslexie zou kunnen zijn (Nicolson & Fawcett, 2007; 2011; Ullman, 2004; Ullman, Sayako, Earle, Walenski, & Janacsek, 2019). Als statistisch leren een cruciale rol speelt in de verwerving van gesproken en geschreven taal in de normale ontwikkeling, dan volgt hieruit de hypothese dat een stoornis in het statistisch leermechanisme zou kunnen resulteren in problemen met taal-verwerving zoals we die zien bij mensen met dyslexie. In overeenstemming met deze hypothese hebben diverse studies aangetoond dat mensen met dyslexie slechter presteren op statistische leertaken dan leeftijdsgenoten zonder dyslexie (bijvoorbeeld Gabay, Thiessen, & Holt, 2015; Jiménez-Fernández, Vaquero, Jiménez, & Defior, 2011; Lum, Ullman, & Conti-Ramsden, 2013; Pavlidou & Williams, 2014).

Het onderzoek naar de veronderstelde relatie tussen statistisch leren en (gesproken en geschreven) taal kent twee benaderingen: (1) de studie van correlaties tussen individuele verschilllen in statistisch leervermogen en scores op taalmaten, en (2) de studie van groepsverschillen tussen mensen met en zonder dyslexie. Het is belangrijk om hier op te merken dat er ook studies zijn gedaan die nulresultaten hebben gerapporteerd, zowel op het gebied van individuele verschillen (bijvoorbeeld Schmalz, Moll, Mulatti, & Schulte-Körne, 2019; West, Vadillo, Shanks, & Hulme, 2017), als op het gebied van groepsverschillen (bijvoorbeeld Kelly, Griffiths, & Frith, 2002; Kerkhoff, de Bree, de Klerk, &

Wijnen, 2013). Deze verschillende bevindingen (d.w.z. sommige studies rapporteren significante resultaten, andere studies rapporteren nulresultaten) hebben geleid tot kritische vragen (bijvoorbeeld Arnon, 2019b; Kidd, Donnelly, & Christiansen, 2017; Schmalz et al., 2019; West et al., 2017). Een voorbeeld van zo'n vraag is: hoe sterk is de relatie tussen statistisch leren aan de ene kant en gesproken en geschreven taal aan de andere kant eigenlijk? Ook: zijn wij wel goed in staat om deze relatie betrouwbaar te onderzoeken, met name in kinderen? Er is behoefte aan grootschalige studies die de relatie tussen statistisch leren en taal onderzoeken op een betrouwbare manier. Daarnaast is er behoefte aan studies die dit doen in relatie tot ontwikkelingsstoornissen zoals dyslexie, om zo inzicht te verkrijgen in de mogelijke onderliggende oorzaken van hun taalproblemen (Arciuli & Conway, 2018).

In deze dissertatie hebben we geprobeerd om deze lacunes te vullen en bewijsmateriaal te verzamelen over de relatie tussen statistisch leren en de verwerving van gesproken en geschreven taal. Wij onderzochten de bijdrage van statistisch leren aan grammatica en lees- en spelvaardigheid in Nederlandse kinderen met en zonder een dyslexiediagnose. Verder hebben we gecontroleerd voor individuele variatie in cognitieve vaardigheden die gerelateerd zijn aan statistisch leren en/of aan de verwerving van gesproken en geschreven taal (bijvoorbeeld aandacht, geheugen, fonologische verwerking; zie e.g. Arciuli, 2017). Bovendien zijn in het kader van deze dissertatie nieuwe taken ontwikkeld die als doel hadden om op een betrouwbare manier het statistischleervermogen van schoolgaande kinderen te meten. In de volgende paragrafen behandelen we de toegepaste methodologieën en verkregen resultaten uit hoofdstukken 2 tot en met 6, gevolgd door de conclusies en implicaties voor toekomstig onderzoek.

Als eerste stelden we onszelf als doel om eerdere statistische leertaken te verbeteren en aan te passen voor gebruik met kinderen. In hoofdstuk 2 ontwikkelden we een nieuwe, "self-paced" visuele statistischleertaak (VSL) en namen we deze af bij 53 kinderen tussen de 5 en 8 jaar (van Witteloostuijn, Lammertink, Boersma, Wijnen, & Rispens, 2019). Eerdere VSL-studies hebben (overwegend) gebruik gemaakt van de zogeheten "two-alternative forced choice" (2-AFC) maat, die het statistischleervermogen meet door middel van expliciete vragen *na* blootstelling aan de VSL-structuur. Dit type leermaten, dat afgenomen wordt nadat het statistischleerproces is voltooid, wordt "offline" genoemd. De VSL-taak die wij in hoofdstuk 2 hebben geïntroduceerd maakt gebruik van een nieuwe leermaat op basis van reactietijden (RTs). De taak is gebaseerd op een

eerdere studie met volwassenen (Siegelman, Bogaerts, Christiansen, & Frost, 2018). Het is belangrijk om te vermelden dat deze RTs verzameld worden gedurende de blootstelling aan de statistische structuur in de VSL-taak (d.w.z. dit is een "online" leermaat). Het ontwikkelen van online maten is essentieel, omdat de betrouwbaarheid van de traditionele 2-AFC taken in twijfel wordt getrokken (bijvoorbeeld Siegelman & Frost, 2015) en omdat offline maten ons niet informeren over het leerproces gedurende de blootstelling aan de statistische structuur (bijvoorbeeld Siegelman, Bogaerts, & Frost, 2017). De studie die gerapporteerd wordt in hoofdstuk 2 was de eerste die aantoont dat een dergelijke online RT-maat in de VSL-taak bruikbaar is onder jonge schoolgaande kinderen: de RT-maat detecteerde dus evidentie van leren. De offline testfase, die achteraf werd afgenomen, toonde geen bewijs dat 5 tot 8 jaar oude kinderen gevoelig waren voor de statistische structuur in de VSL-taak. Kortom, hoofdstuk 2 benadrukt de potentie van online maten van statistisch leren, ook voor gebruik met kinderen.

Hoofdstukken 3 en 4 bestudeerden het statistischleervermogen van individuen met dyslexie: waar hoofdstuk 3 een meta-analyse bevat over voorgaande studies die het visuele "artificial grammar learning" (AGL) paradigma toepassen (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017), bevat hoofdstuk 4 een experimentele studie waarin gebruik werd gemaakt van vernieuwde statistischleermaten (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019). In onze meta-analyse van 13 eerdere studies die de prestaties van individuen met dyslexie onderzochten door middel van de visuele AGL, vonden we een effect van groep op prestaties: individuen met dyslexie presteerden slechter dan leeftijdsgenoten. We vonden ook bewijs voor een publicatie-bias; het lijkt waarschijnlijk dat statistisch significante studies makkelijker gepubliceerd worden dan studies die geen statistisch significant resultaat rapporteren. Zodra we voor deze publicatiebias controleerden, was het effect van groep op de AGL prestaties niet langer significant. Oftewel, onze initiële bevinding is waarschijnlijk iets te optimistisch: het gerapporteerde verschil in visuele AGL prestaties tussen individuen met en zonder dyslexie zou door het in acht nemen van ongepubliceerde studies mogelijk verdwijnen.

In hoofdstuk 4 gebruikten we drie andere paradigma's die statistisch leren meten, om verder inzicht te verkrijgen in de problemen die mensen met dyslexie mogelijk hebben op dit gebied. Honderd kinderen namen deel aan deze

studie: 50 kinderen tussen de 8 en 11 jaar oud met een dyslexiediagnose en 50 kinderen zonder dyslexie van dezelfde leeftijd. Naast de nieuwe self-paced VSL-taak (van Witteloostuijn, Lammertink et al., 2019, Chapter 2), gebruikten we een nieuwe online maat van leren in een zogeten auditieve "nonadjacent dependency learning" (A-NADL) taak (Lammertink, van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019) en een meer traditionele visueel-motorische "serial reaction time" (SRT) taak. Door deze drie paradigma's te gebruiken, wilden we statistisch leren zo breed mogelijk testen. Niet alleen verschillen de drie taken op het gebied van domein (visueel, auditief en visueel-motorisch), ze verschillen ook qua type statistische structuur (bijvoorbeeld adjecent en nonadjecent, oftewel naburig en meer afgelegen, structuren). Daarnaast controleerden we in onze analyses voor individuele verschillen in volgehouden aandacht en in kortetermijn- en werkgeheugen, omdat deze cognitieve vaardigheden verband houden met prestaties op statistische leertaken (Arciuli, 2017; Baker, Olson, & Behrmann, 2004). Bovendien is bekend dat individuen met dyslexie moeilijkheden ervaren op het gebied van aandacht en geheugen (Bosse, Tainturier, & Valdois, 2007; Cowan et al., 2017). Onze bevindingen op de drie statistischleermaten waren consistent: we vonden bewijs dat zowel kinderen met als zonder dyslexie gevoelig zijn voor de statistische structuren die we ze aanboden. De andere consistente bevinding was dat we geen bewijs vonden voor (of tegen) een verschil in prestaties tussen kinderen met en zonder dyslexie op de VSL-, A-NADL- en/of de SRT-taak. Kortom, de resultaten van hoofdstuk 4 (en 3) ondersteunen niet de hypothese dat een domeinoverstijgend statistischleerprobleem de (gesproken- en geschreven-) taalproblemen van individuen met dyslexie veroorzaakt.

Hoewel we geen bewijs vonden voor (of tegen) een verschil in statistisch-leervermogen tussen kinderen met en zonder dyslexie, is het nog steeds mogelijk dat individuele verschillen in statistisch leren bijdragen aan verschillen in gesproken en geschreven taal. Deze hypothese komt voort uit twee overwegingen: (1) hoofdstuk 4 toonde aan dat zowel kinderen met als zonder dyslexie leerden in alle drie de taken die statistisch leren meten, en (2) eerder onderzoek heeft laten zien dat de relatie tussen statistisch leren en taal ook bestaat in zich normaal ontwikkelde populaties. In de hoofdstukken 5 en 6 bestudeerden we de individuele verschillen in statistisch leren in onze steekproef van honderd kinderen met en zonder dyslexie. We onderzochten of deze individuele verschillen geassocieerd waren met scores op taken die grammaticale vaardigheden toetsen (inflectionele morfologie en syntaxis; van Witteloostuijn, Boersma, Wijnen, &

Rispens, submitted, see chapter 6) en met taken die verband houden met geschreven taal (technisch lezen en spellen; van Witteloostuijn, Boersma, Wijnen, & Rispens, under review, see chapter 5). Daarnaast berekenden we de interne consistentie en betrouwbaarheid van onze (nieuwe) statistischleermaten door middel van de zogeten "split-half betrouwbaarheid" (zie ook bijvoorbeeld Arnon, 2019a; Siegelman, Bogaerts, Christiansen, & Frost, 2017b) en controleerden we opnieuw voor een reeks variabelen op het niveau van de proefpersoon (bijvoorbeeld leeftijd, aandacht, geheugen, fonologische verwerking). Net als besproken voor hoofdstukken 3 en 4 waren de bevindingen in hoofdstukken 5 en 6 consistent: we vinden geen bewijs voor (of tegen) de veronderstelde relatie tussen statistisch leren aan de ene kant en prestaties op grammaticale taken (inflectionele morfologie en syntaxis) en geschreven taal (technisch lezen en spellen) aan de andere kant. De split-half betrouwbaarheid van de statistischleermaten gebruikt in deze dissertatie benaderden de norm voor gestandaardiseerde psychometrische tests van 0,80 en varieerden tussen 0,70 op de VSL offline maat en 0,79 op de A-NADL online maat. Exploratieve analyses bevestigden dat er een relatie is tussen fonologische vaardigheden (bijvoorbeeld fonologische verwerking en fonologisch kortetermijn- en werkgeheugen) en zowel grammaticale als lees- en spelvaardigheden in kinderen met en zonder dyslexie. Tot slot onderstreepte hoofdstuk 6 de eerdere bevinding dat individuen met dyslexie subtiele problemen ervaren op het gebied van grammatica. Deze problemen vonden we niet alleen op het gebied van inflectionele morfologie (bijvoorbeeld onregelmatige meervoudsvorming en verledentijdsvorming), maar ook op het gebied van syntaxis (bijvoorbeeld werkwoordsvolgorde).

Als we alle resultaten gepresenteerd in deze dissertatie samennemen, voorzien zij ons niet van bewijs voor (of tegen) een link tussen een domeinoverstijgend statistischleermechanisme en de verwerving van gesproken en geschreven taal. Deze resultaten kwamen voort uit twee verschillende onderzoekslijnen: (1) we vonden geen bewijs dat individuele verschillen in statistisch leren bijdragen aan de prestaties op taken die het (gesproken en geschreven) taalvermogen testen in kinderen met en zonder dyslexie, en (2) hoewel kinderen gevoelig waren voor de statistische patronen die we ze aanboden in de drie statistische leertaken, vonden we geen bewijs voor (of tegen) een verschil in prestaties tussen kinderen met en zonder dyslexie. We kunnen dus niet uitsluiten dat de relatie tussen statistisch leren en taal minder sterk is dan van tevoren

verondersteld, waardoor de associatie mogelijk alleen optreedt wanneer specifieke methodologische keuzes gemaakt worden. Verder lijkt het op basis van onze bevindingen onwaarschijnlijk dat individuen met dyslexie een domeinoverstijgend en uitgebreid probleem hebben met statistisch leren. Meer onderzoek in de vorm van grootschalige en pregeregistreerde studies, evenals meta-analyses, is nodig voordat we definitieve conclusies kunnen trekken wat betreft de bijdrage van (domeinoverstijgend) statistischleervermogen aan de verwerving van gesproken en geschreven taal, zowel in zich normaal ontwikkelende kinderen als in kinderen met ontwikkelingsstoornissen zoals dyslexie.

About the author

Merel van Witteloostuijn was born on January 1st 1989 in Maastricht, The Netherlands. After graduating from the Mollerlyceum in 2008, she started a bachelor in Linguistics at Utrecht University with a minor in cognition at the Psychology department. Already during her bachelor degree, she discovered her interest in psycholinguistics and language acquisition, which drove her to continue with the research master Linguistics at Utrecht University in 2011. This research master allowed her to engage in several research projects in the area of developmental disorders: at the University of Groningen, she was involved in a project that investigated the language acquisition and cognitive development of children with autism spectrum disorders, while at the Unversity of Amsterdam, she worked in a project that examined grammar and pragmatics in children with autism spectrum disorders and children with developmental language disorders. Following her cum laude graduation from the research master in 2013, she continued to work on the latter project as a research assistant, followed by another research assistant position at Utrecht University in a project that looked at children's texting behavior and its impact on their cognitive and social development.

In May 2015, Merel began her PhD at the University of Amsterdam, working on the project "Examining the contribution of procedural learning to grammar and literacy acquisition in children" (financed through a VIDI-grant awarded to prof. dr. Judith Rispens by the Dutch Research Council). During her PhD years, she spent three months as a guest researcher at the Hebrew University of Jerusalem, published eight articles in international journals, and presented her work at national and international conferences. Moreover, she was active in the PhD council of the Faculty of Humanities, taught the course "Second Language Acquisition", gave guest lectures, and supervised students working on their internships and theses. In January 2020, Merel started working as a data analyst at the Dutch Ministry of Education, Culture and Science.